Advances in Intelligent Systems and Computing 240

Jerzy Świątek
Adam Grzech
Paweł Świątek
Jakub M. Tomczak *Editors*

Advances in Systems Science

Proceedings of the International Conference on Systems Science 2013 (ICSS 2013)

Springer

Advances in Intelligent Systems and Computing

Volume 240

Series Editor

Janusz Kacprzyk, Warsaw, Poland

For further volumes:
http://www.springer.com/series/11156

Jerzy Świątek · Adam Grzech
Paweł Świątek · Jakub M. Tomczak
Editors

Advances in Systems Science

Proceedings of the International Conference on Systems Science 2013 (ICSS 2013)

Springer

Editors

Jerzy Świątek
Institute of Informatics
Wroclaw University of Technology
Wrocław
Poland

Adam Grzech
Institute of Informatics
Wroclaw University of Technology
Wrocław
Poland

Paweł Świątek
Institute of Informatics
Wroclaw University of Technology
Wrocław
Poland

Jakub M. Tomczak
Institute of Informatics
Wroclaw University of Technology
Wrocław
Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-01856-0 ISBN 978-3-319-01857-7 (eBook)
DOI 10.1007/978-3-319-01857-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013946038

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The International Conference on Systems Science 2013 (ICSS 2013) was the 18th event of the series of international scientific conferences for researchers and practitioners in the fields of systems science and systems engineering. The conference took place in Wroclaw, Poland during September 10–12, 2013 and was organized by Wroclaw University of Technology and co-organized by: Committee of Automatics and Robotics of Polish Academy of Sciences, Committee of Computer Science of Polish Academy of Sciences and Polish Section of IEEE.

The first International Conference on Systems Science organized by Wroclaw University of Technology was held in 1974 and was organized every year till 1980 when Coventry Polytechnic, Coventry, UK started to co-organize parallel scientific events called International Conference on Systems Engineering (ICSE) every two years. In 1984 the Wright State University, Dayton, Ohio, USA joined to co-operate and organize International Conference on Systems Engineering. In 1990 the ICSE moved from Wright State University, Dayton, Ohio, USA to Nevada State University, Las Vegas, USA. Now, the International Conference on Systems Science is organized every three years in Wroclaw, by Wroclaw University of Technology and in the remaining years the International Conference on Systems Engineering is organized in Las Vegas by Nevada State University or in Coventry by Coventry University. The aim of the International Conference on Systems Science (ICSS) and the International Conference on Systems Engineering (ICSE) series was to provide an international forum for scientific research in systems science and systems engineering.

This year, we received almost 140 papers from 34 countries. Each paper was reviewed by at least two members of Program Committee or Board of Reviewers. Only 76 best papers were selected for oral presentation and publication in the International Conference on Systems Science 2013 proceedings. The final acceptance rate was 55%.

The papers included in the proceedings cover the following topics:

- Control Theory
- Databases and Data Mining
- Image and Signal Processing

- Machine Learning
- Modelling and Simulation
- Operational Research
- Service Science
- Time Series and System Identification

Accepted and presented papers highlight new trends and challenges in systems science and systems engineering. The presenters show how new research could lead to new and innovative applications. We do hope you will find these results useful and inspiring for your future research.

We would like to thank the Program Committee and Board of Reviewers, essential for reviewing the papers to ensure a high standard.

Jerzy Świątek

Organization

Program Committee

P. Albertos (Spain)
A.V. Balakrishnan (USA)
S. Bańka (Poland)
A. Bartoszewicz (Poland)
L. Borzemski (Poland)
V.N. Burkov (Russia)
K.J. Burnham (UK)
L.M. Camarinha-Matos (Portugal)
A. Grzech (Poland)
K. Hasegawa (Japan)
T. Hasegawa (Japan)
A. Isidori (Italy)
D.J.G. James (UK)
J. Józefczyk (Poland)
J. Kacprzyk (Poland)
T. Kaczorek (Poland)
R. Kaszyński (Poland)
L. Keviczky (Hungary)
G. Klir (USA)
P. Kontogiorgis (USA)
J. Korbicz (Poland)
G.L. Kovacs (Hungary)
V. Kucera (Czech Republic)
A.B. Kurzhanski (Russia)
H. Kwaśnicka (Poland)
N. Lavesson (Sweden)
B. Neumann (Germany)
J.J. Lee (Korea)
S.Y. Nof (USA)
W. Pedrycz (Canada)
F. Pichler (Austria)
G.P. Rao (India)
A. Rindos (USA)
L. Rutkowski (Poland)
E. Szczerbicki (Australia)
H. Selvaraj (USA)
R. Słowiński (Poland)
M. Sugisaka (Japan)
A. Sydow (Germany)
J. Świątek (Poland)
R. Tadeusiewicz (Poland)
Y. Takahara (Japan)
M. Thoma (Germany)
S.G. Tzafestas (Greece)
H. Unbehauen (Germany)
R. Vallee (France)
J. Węglarz (Poland)
L.A. Zadeh (USA)

Organizing Committee

Conference Chairman
Jerzy Świątek

Conference Co-Chairman
Adam Grzech

Financial Chairman
Paweł Świątek

Technical Chairmen
Krzysztof Brzostowski
Jarosław Drapała

Conference Secretary
Jakub M. Tomczak

Contents

Keynote Speakers

Control Theory

Databases and Data Mining

Image and Signal Processing

Machine Learning

Modelling and Simulation

Operational Research

Service Science

Time Series and System Identification

Erratum

Decoupling Zeros of Positive Continuous-Time Linear Systems and Electrical Circuit

Tadeusz Kaczorek

Białystok University of Technology, Faculty of Electrical Engineering,
Wiejska 45D, 15-351 Bialystok Poland
kaczorek@isep.pw.edu.pl

Abstract. Necessary and sufficient conditions for the reachability and observability of the positive continuous-time linear systems are established. Definitions of the input-decoupling zeros, output-decoupling zeros and input-output decoupling zeros are proposed. Some properties of the decoupling zeros are discussed. Decoupling zeros of positive electrical circuits are also addressed.

Keywords: decoupling zeros, positive, continuous-time, linear system, positive electrical circuits, observability, reachability.

1 Introduction

In positive systems inputs, state variables and outputs take only non-negative values. Examples of positive systems are industrial processes involving chemical reactors, heat exchangers and distillation columns, storage systems, compartmental systems, water and atmospheric pollution models. A variety of models having positive linear behavior can be found in engineering, management science, economics, social sciences, biology and medicine, etc. An overview of state of the art in positive linear theory is given in the monographs [2, 3].

The notions of controllability and observability and the decomposition of linear systems have been introduced by Kalman [15, 16]. Those notions are the basic concepts of the modern control theory [1, 2, 4, 14, 17, 21]. They have been also extended to positive linear systems [2, 3].

The reachability and controllability to zero of standard and positive fractional discrete-time linear systems have been investigated in [9] and controllability and observability of electrical circuits in [6, 8, 10]. The decomposition of positive discrete-time linear systems has been addressed in [5]. The notion of decoupling zeros of standard linear systems have been introduced by Rosenbrock [17]. The zeros of linear standard discrete-time system have been addressed in [20] and zeros of positive continuous-time and discrete-time linear systems has been defined in [18, 19]. The decoupling zeros of positive discrete-time linear systems has been introduced in [7] and of positive continuous-time systems in [12, 13]. The positivity and reachability of fractional electrical circuits have been investigated in [8].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_1, © Springer International Publishing Switzerland 2014

In this paper the notions of decoupling zeros will be extended for positive continuous-time linear systems and electrical circuits.

The paper is organized as follows. In section 2 the basic definitions and theorems concerning reachability and observability of positive continuous-time linear systems are given. The decomposition of the pair (*A*,*B*) and (*A*,*C*) of positive linear system is addressed in section 3. The main result of the paper is given in section 4 where the definitions of the decoupling-zeros are proposed. The positive electrical circuits are addressed in section 5 and decoupling zeros of positive electrical circuits in section 6. Concluding remarks are given in section 7.

The following notation will be used: $\Re$ - the set of real numbers, $\Re^{n\times m}$ - the set of $n\times m$ real matrices, $\Re_+^{n\times m}$ - the set of $n\times m$ matrices with nonnegative entries and $\Re_+^n = \Re_+^{n\times 1}$, M_n - the set of $n\times n$ Metzler matrices (real matrices with nonnegative off-diagonal entries), I_n - the $n\times n$ identity matrix.

2 Reachability and Observability of Positive Continuous-Time Linear Systems

2.1 Reachability of Positive Systems

Consider the linear continuous-time system

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t)\\ y(t) &= Cx(t) + Du(t)\end{aligned} \tag{2.1}$$

where $x(t)\in\Re^n$, $u(t)\in\Re^m$, $y(t)\in\Re^p$ are the state, input and output vectors and $A\in\Re^{n\times n}$, $B\in\Re^{n\times m}$, $C\in\Re^{p\times n}$, $D\in\Re^{p\times m}$.

Definition 2.1. [2, 3] The linear system (2.1) is called (internally) positive if $x(t)\in\Re_+^n$ and $y(t)\in\Re_+^p$, $t\ge 0$ for any $x(0)=x_0\in\Re_+^n$ and every $u(t)\in\Re_+^m$, $t\ge 0$.

Theorem 2.1. [2, 3] The system (2.1) is positive if and only if

$$A\in M_n,\quad B\in\Re_+^{n\times m},\quad C\in\Re_+^{p\times n},\quad D\in\Re_+^{p\times m} \tag{2.2}$$

Definition 2.2. The positive system (2.1) (or positive pair (*A*,*B*)) is called reachable at time t_f if for any given final state $x_f\in\Re_+^n$ there exists an input sequence $u(t)\in\Re_+^m$, $t\in[0,t_f]$ which steers the state of the system from zero state ($x(0)=0$) to state $x_f\in\Re_+^n$, i.e. $x(t_f)=x_f$.

A column $a \in \Re_+^n$ (row $a^T \in \Re_+^n$) is called monomial if only one its entry is positive and the remaining entries are zero. A real matrix $A \in \Re_+^{n \times n}$ is called monomial if each its row and each its column contains only one positive entry and the remaining entries are zero.

Theorem 2.2. The positive system (2.1) is reachable at time $t \in [0, t_f]$ if and only if the matrix $A \in M_n$ is diagonal and the matrix $B \in \Re_+^{n \times n}$ is monomial.
Proof is given in [12].

2.2 Observability of Positive Systems

Consider the positive system

$$\dot{x}(t) = Ax(t) \tag{2.3a}$$

$$y(t) = Cx(t) \tag{2.3b}$$

where $x(t) \in \Re_+^n$, $y(t) \in \Re_+^p$ and $A \in M_n$, $C \in \Re_+^{p \times n}$.

Definition 2.3. The positive system (2.3) is called observable if knowing the output $y(t) \in \Re_+^p$ and its derivatives $y^{(k)}(t) = \dfrac{d^k y(t)}{dt^k} \in \Re_+^p$, k = 1,2,…,n – 1 it is possible to find the initial values $x_0 = x(0) \in \Re_+^n$ of $x(t) \in \Re_+^n$.

Theorem 2.3. The positive system (2.3) is observable if and only if the matrix $A \in M_n$ is diagonal and the matrix

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{2.4}$$

has n linearly independent monomial rows.
Proof is given in [12].

3 Decomposition of the Pairs (*A*,*B*) and (*A*,*C*)

3.1 Decomposition of the Pair (*A*,*B*)

Consider the pair (A,B) with A being diagonal

$$A = \operatorname{diag}[a_{11}, a_{22}, \ldots, a_{n,n}] \in M_n \tag{3.1a}$$

and the matrix B with m linearly independent columns $B_1, B_2, ..., B_m$

$$B = [B_1 \quad B_2 \quad ... \quad B_m]. \tag{3.1b}$$

By Theorem 2.2 the pair (3.1) is unreachable if $m < n$.

It will be shown that in this case the pair can be decomposed into the reachable pair $(\overline{A}_1, \overline{B}_1)$ and unreachable pair $(\overline{A}_2, \overline{B}_2 = 0)$.

Theorem 3.1. For the unreachable pair (3.1) ($m < n$) there exists a monomial matrix $P \in \Re_+^{n\times n}$ such that the pair (A,B) can be reduced to the form

$$\overline{A} = PAP^{-1} = \begin{bmatrix} \overline{A}_1 & 0 \\ 0 & \overline{A}_2 \end{bmatrix}, \quad \overline{B} = PB = \begin{bmatrix} \overline{B}_1 \\ 0 \end{bmatrix} \tag{3.2}$$

where $\overline{A}_1 = \text{diag}\,[\overline{a}_{11}, \overline{a}_{22}, ..., \overline{a}_{n_1,n_1}] \in M_{n_1}$, $\overline{A}_2 = \text{diag}\,[\overline{a}_{n_1+1,n_1+1}, ..., \overline{a}_{n,n}] \in M_{n_2}$, $\overline{B}_1 \in \Re_+^{n_1\times m}$, $n = n_1 + n_2$, the pair $(\overline{A}_1, \overline{B}_1)$ is reachable and the pair $(\overline{A}_2, \overline{B}_2 = 0)$ is unreachable.

Proof. Performing on the matrix B the following elementary row operations:

- interchange the i-th and j-th rows, denoted by $L[i, j]$,
- multiplication of i-th rows by positive number c, denoted by $L[i \times c]$,

we may reduce the matrix B to the form $\begin{bmatrix} \overline{B}_1 \\ 0 \end{bmatrix}$, where $\overline{B}_1 \in \Re_+^{n_1\times m}$ is monomial with positive entries equal to 1. Performing the same elementary row operations on the identity matrix In we obtain the desired monomial matrix P. It is well-known [3] that $P^{-1} \in \Re_+^{n\times n}$ and for diagonal matrix A we have $\overline{A} = PAP^{-1} = \begin{bmatrix} \overline{A}_1 & 0 \\ 0 & \overline{A}_2 \end{bmatrix}$. □

Example 3.1. Consider the unreachable pair (3.1) with

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 3 \\ 0 & 0 \\ 2 & 0 \end{bmatrix}. \tag{3.3}$$

Performing on the matrix B the following elementary row operations $L[1,3]$, $L[1\times 1/2]$, $L[2,3]$, $L[2\times 1/3]$ we obtain

$$\overline{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{3.4}$$

Performing the same elementary row operations on the identity matrix I_3 we obtain the desired monomial matrix

$$P = \begin{bmatrix} 0 & 0 & 1/2 \\ 1/3 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{3.5}$$

and

$$PB = \begin{bmatrix} 0 & 0 & 1/2 \\ 1/3 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 0 & 0 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \overline{B} = \begin{bmatrix} \overline{B}_1 \\ 0 \end{bmatrix},$$
$$\overline{A} = PAP^{-1} = \begin{bmatrix} 0 & 0 & 1/2 \\ 1/3 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 3 & 0 \\ 0 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} = \begin{bmatrix} \overline{A}_1 & 0 \\ 0 & \overline{A}_2 \end{bmatrix}. \tag{3.6}$$

The positive pair

$$\overline{A}_1 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \overline{B}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.7}$$

is reachable and the pair $(\overline{A}_2, 0)$ is unreachable.

3.2 Decomposition of the Pair (*A*,*C*)

Let the observability matrix

$$O_n = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \Re_+^{pn\times n} \tag{3.8}$$

of the positive unobservable system has $n_1 < n$ linearly independent monomial rows.

If the conditions

$$Q_k A Q_j^T = 0 \text{ for } k = 1,2,\dots,n_1 \text{ and } j = n_1 + 1,\dots,n \tag{3.9}$$

are satisfied then there exists the monomial matrix [5, 6]

$$Q^T = [Q_{j_1}{}^T \ \dots \ Q_{j_1\overline{d}_1}{}^T \ \ Q_{j_2}^T \ \dots \ Q_{j_2\overline{d}_2}^T \ \dots \ Q_{j_l\overline{d}_l}^T \ \ Q_{n_1+1}^T \ \dots \ Q_n^T]^T \in \Re_+^{n\times n} \tag{3.10a}$$

where

$$Q_{j_1} = C_{j_1},\dots,Q_{j_1\overline{d}_1} = C_{j_1}A^{\overline{d}_1-1}, Q_{j_2} = C_{j_2},\dots,Q_{j_2\overline{d}_2} = C_{j_2}A^{\overline{d}_2-1},\dots,Q_{j_l\overline{d}_l} = C_{j_l}A^{\overline{d}_l-1} \tag{3.10b}$$

and $\overline{d}_j$, $j = 1,\dots,l$ are some natural numbers.

Theorem 3.2. Let the positive system (2.3) be unobservable and let there exist the monomial matrix (3.10). Then the pair (*A*,*C*) of the system can be reduced by the use of the matrix (3.10) to the form

$$\hat{A} = QAQ^{-1} = \begin{bmatrix} \hat{A}_1 & 0 \\ 0 & \hat{A}_2 \end{bmatrix}, \quad \hat{C} = CQ^{-1} = [\hat{C}_1 \quad 0]$$
$$\hat{A}_1 \in \Re_+^{n_1 \times n_1}, \quad \hat{A}_2 \in \Re_+^{n_2 \times n_2}, \quad (n_2 = n - n_1), \quad \hat{C}_1 \in \Re_+^{p \times n_1} \tag{3.11}$$

where the pair $(\hat{A}_1, \hat{C}_1)$ is observable and the pair $(\hat{A}_2, \hat{C}_2 = 0)$ is unobservable.

Proof is given in [5].

Example 3.2. Consider the unobservable pair

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad C = [0 \quad 0 \quad 1]. \tag{3.12}$$

In this case the observability matrix

$$Q_3 = \begin{bmatrix} C \\ CA \\ CA^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.13}$$

has only one monomial row $Q_1 = C$, i.e. $n_1 = 1$ and the conditions (3.9) are satisfied for $Q_2 = [1 \quad 0 \quad 0]$ and $Q_3 = [0 \quad 1 \quad 0]$ since $Q_1 A Q_j^T = 0$ for $j = 2,3$. The matrix (3.10) has the form

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{3.14}$$

Using (3.11) and (3.14) we obtain

$$\hat{A} = QAQ^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} = \begin{bmatrix} \hat{A}_1 & 0 \\ 0 & \hat{A}_2 \end{bmatrix},$$
$$\hat{C} = CQ^{-1} = [0 \quad 0 \quad 1] \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = [1 \quad 0 \quad 0] = [\hat{C}_1 \quad 0], \tag{3.15}$$

where

$$\hat{A}_1 = [-1],\ \hat{A}_2 = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix},\ \hat{C}_1 = [1]. \tag{3.16}$$

The pair $(\hat{A}_1, \hat{C}_1)$ is observable and the pair $(\hat{A}_2, 0)$ is unobservable.

4 Decoupling Zeros of the Positive Systems

It is well-known [17] that for standard linear systems the input-decoupling zeros are the eigenvalues of the matrix $\overline{A}_2$ of the unreachable (uncontrollable) part $(\overline{A}_2, \overline{B}_2 = 0)$.

In a similar way we will define the input-decoupling zeros of the positive continuous-time linear systems.

Definition 4.1. Let $\overline{A}_2$ be the matrix of unreachable part of the system (2.1). The zeros $s_{i1}, s_{i2}, ..., s_{i\overline{n}_2}$ of the characteristic polynomial

$$\det[I_{\overline{n}_2} s - \overline{A}_2] = s^{\overline{n}_2} + \overline{a}_{\overline{n}_2 - 1} s^{\overline{n}_2 - 1} + ... + \overline{a}_1 s + \overline{a}_0 \tag{4.1}$$

of the matrix $\overline{A}_2$ are called the input-decoupling zeros of the positive system (2.1). The list of the input-decoupling zeros will be denoted by $Z_i = \{s_{i1}, s_{i2}, ..., s_{i\overline{n}_2}\}$.

Theorem 4.1. The state vector $x(t)$ of the positive system (2.1) is independent of the input-decoupling zeros for any input $u(t)$ and zero initial conditions.

Proof. From (2.1) for zero initial conditions $x(0) = 0$ we have

$$X(s) = [I_n s - A]^{-1} BU(s), \tag{4.2}$$

where $X(s)$ and $U(s)$ are Laplace transforms of $x(t)$ and $u(t)$, respectively. Taking into account (3.2) we obtain

$$\begin{aligned} X(s) &= [I_n s - P^{-1}\overline{A}P]^{-1} P^{-1} BU(s) = P^{-1}[I_n s - \overline{A}]^{-1} \overline{B} U(s) \\ &= P^{-1} \begin{bmatrix} I_{\overline{n}_1} s - \overline{A}_1 & 0 \\ 0 & I_{\overline{n}_2} s - \overline{A}_2 \end{bmatrix}^{-1} \begin{bmatrix} \overline{B}_1 \\ 0 \end{bmatrix} U(s) = P^{-1} \begin{bmatrix} [I_{\overline{n}_1} s - \overline{A}_1]^{-1} \overline{B}_1 \\ 0 \end{bmatrix} U(s). \end{aligned} \tag{4.3}$$

From (4.3) it follows that $X(s)$ is independent of the matrix $\overline{A}_2$ and of the input-decoupling zeros for any input $u(t)$. □

Example 4.1. (continuation of Example 3.1) In Example 3.1 it was shown that for the unreachable pair (3.1) the matrix $\overline{A}_2$ has the form $\overline{A}_2 = [-2]$. Therefore, the positive system (3.1) with (3.3) has one input-decoupling zero $s_{i1} = -2$.

For standard continuous-time linear systems the output-decoupling zeros are defined as the eigenvalues of the matrix of the unobservable part of the system. In a similar way we will define the output-decoupling zeros of the positive continuous-time linear systems.

Definition 4.2. Let $\hat{A}_2$ be the matrix of unobservable part of the system (2.3). The zeros $s_{o1}, s_{o2}, ..., s_{o\hat{n}_2}$ of the characteristic polynomial

$$\det[I_{\hat{n}_2} s - \hat{A}_2] = s^{\hat{n}_2} + \hat{a}_{\hat{n}_2 - 1} s^{\hat{n}_2 - 1} + ... + \hat{a}_1 s + \hat{a}_0 \tag{4.4}$$

of the matrix $\hat{A}_2$ are called the output-decoupling zero of the positive system (2.3).

The list of the output-decoupling zeros will be denoted by $Z_o = \{s_{o1}, s_{o2}, ..., s_{o\hat{n}_2}\}$.

Theorem 4.2. The output vector $y(t)$ of the positive system (2.3) is independent of the output-decoupling zeros for any input $\overline{u}(t) = Bu(t)$ and zero initial conditions. Proof is similar to the proof of Theorem 4.1.

Example 4.2. (continuation of Example 3.2) In Example 3.2 it was shown that the matrix $\hat{A}_2$ of the positive unobservable pair (3.12) has the form

$$\hat{A}_2 = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \tag{4.5}$$

and the positive system has two output-decoupling zero $s_{o1} = -1$, $s_{o2} = -2$.

Following the same way as for standard continuous-time linear systems we define the input-output decoupling zeros of the positive systems as follows.

Definition 4.3. Zeros $s_{io}^{(1)}, s_{io}^{(2)}, ..., s_{io}^{(k)}$ which are simultaneously the input-decoupling zeros and the output-decoupling zeros of the positive system are called the input-output decoupling zeros of the positive system, i.e.

$$s_{io}^{(j)} \in Z_i \text{ and } s_{io}^{(j)} \in Z_o \text{ for } j = 1,2,\dots,k;\ k \le \min(\overline{n}_2, \hat{n}_2). \tag{4.6}$$

The list of input-output decoupling zeros will be denoted by $Z_{io} = \{z_{io}^{(1)}, z_{io}^{(2)}, ..., z_{io}^{(k)}\}$.

Example 4.3. Consider the positive system with the matrices *A*, *B*, *C* given by (3.3) and (3.12). In Example 4.1 it was shown that the positive system has one input-decoupling zero $s_{i1} = -2$ and in Example 4.2 that the system has two output-decoupling zeros $s_{o1} = -1$, $s_{o2} = -2$. Therefore, by Definition 4.3 the positive system has one input-output decoupling zero $s_{io}^{(1)} = -2$.

5 Positive Electrical Circuits

Example 5.1. Consider the electrical circuit shown in Figure 1 with given conductances $G_1, G'_1, G_2, G'_2, G_{12}$, capacitances C_1, C_2 and source voltages e_1, e_2.

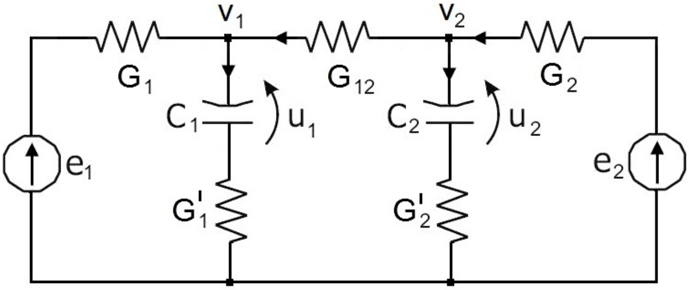


Fig. 1. Electrical circuit

Using the Kirchhoff's laws we can write the equations

$$\frac{d}{dt}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} \dfrac{G'_1}{C_1} & 0 \\ 0 & \dfrac{G'_2}{C_2} \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} - \begin{bmatrix} \dfrac{G'_1}{C_1} & 0 \\ 0 & \dfrac{G'_2}{C_2} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{5.1}$$

and

$$\begin{bmatrix} -G_{11} & G_{12} \\ G_{12} & -G_{22} \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = -\begin{bmatrix} G'_1 & 0 \\ 0 & G'_2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} - \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{5.2a}$$

where

$$G_{11} = G_1 + G'_1 + G_{12}, \quad G_{22} = G_2 + G'_2 + G_{12}. \tag{5.2b}$$

Taking into account that the matrix

$$\begin{bmatrix} -G_{11} & G_{12} \\ G_{12} & -G_{22} \end{bmatrix} \tag{5.3}$$

is nonsingular and

$$-\begin{bmatrix} -G_{11} & G_{12} \\ G_{12} & -G_{22} \end{bmatrix}^{-1} \in \Re_+^{2\times 2} \tag{5.4}$$

from (5.2a) we obtain

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = -\begin{bmatrix} -G_{11} & G_{12} \\ G_{12} & -G_{22} \end{bmatrix}^{-1}\left\{\begin{bmatrix} G'_1 & 0 \\ 0 & G'_2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}\right\} \tag{5.5}$$

Substitution of (5.5) into (5.1) yields

$$\frac{d}{dt}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = A\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + B\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{5.6}$$

where

$$A = -\begin{bmatrix} \frac{G'_1}{C_1} & 0 \\ 0 & \frac{G'_2}{C_2} \end{bmatrix}\begin{bmatrix} -G_{11} & G_{12} \\ G_{12} & -G_{22} \end{bmatrix}^{-1}\begin{bmatrix} G'_1 & 0 \\ 0 & G'_2 \end{bmatrix} - \begin{bmatrix} \frac{G'_1}{C_1} & 0 \\ 0 & \frac{G'_2}{C_2} \end{bmatrix} \in M_2, \tag{5.7a}$$

$$B = -\begin{bmatrix} \frac{G'_1}{C_1} & 0 \\ 0 & \frac{G'_2}{C_2} \end{bmatrix}\begin{bmatrix} -G_{11} & G_{12} \\ G_{12} & -G_{22} \end{bmatrix}^{-1}\begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix} \in \Re_+^{2\times 2}. \tag{5.7b}$$

From (5.7) it follows that A is Metzler matrix and the matrix B has nonnegative entries. Therefore, the electrical circuit is positive for all values of the conductances and capacitances.

In general case we have the following theorem.

Theorem 5.1. The electrical circuit shown in Figure 2 is positive for all values of the conductances, capacitances and source voltages.
Proof is given in [8, 10].

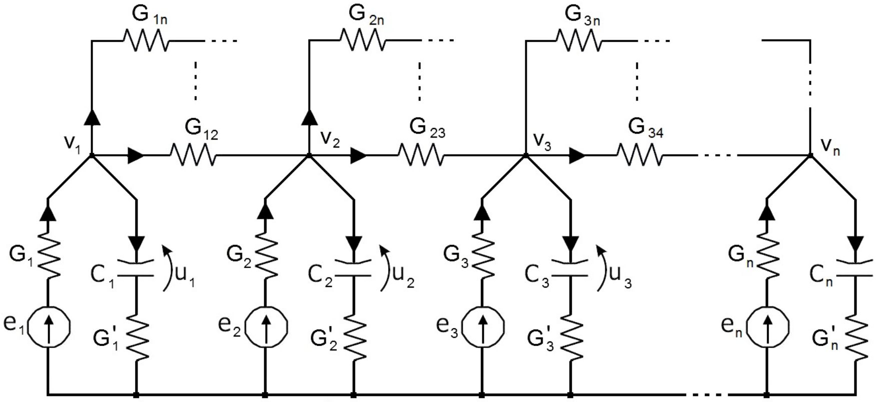


Fig. 2. Electrical circuit

Note that the standard electrical circuit shown in Figure 5.2 is reachable for all nonzero values of the conductances and capacitances since $\det B \neq 0$.

Theorem 5.2. The electrical circuit shown in Figure 5.2 is reachable if and only if

$$G_{k,j} = 0 \quad \text{for} \quad k \neq j \quad \text{and} \quad k, j = 1,...n.. \tag{5.8}$$

Proof. It is easy to see that the matrices $A \in M_n$ and $B \in \Re_+^{n \times n}$ are both diagonal matrices if and only if the condition (5.8) is satisfied. In this case by Theorem 2.2 the electrical circuit is reachable if and only if the conditions (5.8) are met. □

Example 5.2. Consider the electrical circuit shown in Figure 3 with given resistances R_1, R_2, R_3, inductances L_1, L_2, L_3 and source voltages e_1, e_3.

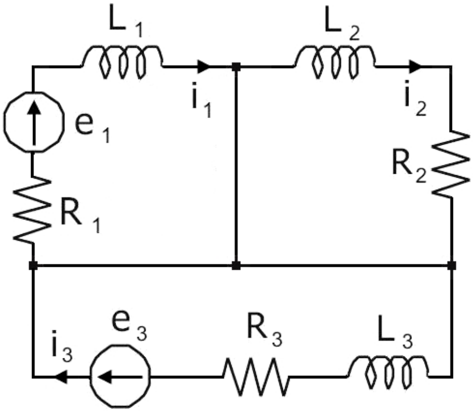


Fig. 3. Electrical circuit

Using the Kirchhoff's laws we can write the equations

$$L_1 \frac{di_1}{dt} = -R_1 i_1 + e_1$$
$$L_2 \frac{di_2}{dt} = -R_2 i_2 \tag{5.9}$$
$$L_3 \frac{di_3}{dt} = -R_3 i_3 + e_3$$

which can be written in the form

$$\frac{d}{dt}\begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} = A \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} + B \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{5.10a}$$

where

$$A = \begin{bmatrix} -\frac{R_1}{L_1} & 0 & 0 \\ 0 & -\frac{R_2}{L_2} & 0 \\ 0 & 0 & -\frac{R_3}{L_3} \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{L_1} & 0 \\ 0 & 0 \\ 0 & \frac{1}{L_3} \end{bmatrix}. \tag{5.10b}$$

By Theorem 2.2 the positive electrical circuit (or the pair (5.10b)) is unreachable since $n = 3 < m = 2$.

The unreachable pair (5.10b) can be decomposed into reachable pair $(\overline{A}_1, \overline{B}_1)$ and unreachable pair $(\overline{A}_2, \overline{B}_2 = 0)$

In this case the monomial matrix P has the form

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{5.11}$$

and we obtain

$$\overline{B} = PB = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{L_1} & 0 \\ 0 & 0 \\ 0 & \frac{1}{L_3} \end{bmatrix} = \begin{bmatrix} \frac{1}{L_1} & 0 \\ 0 & \frac{1}{L_3} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \overline{B}_1 \\ 0 \end{bmatrix}, \quad \overline{B}_1 = \begin{bmatrix} \frac{1}{L_1} & 0 \\ 0 & \frac{1}{L_3} \end{bmatrix},$$

$$\overline{A} = PAP^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -\frac{R_1}{L_1} & 0 & 0 \\ 0 & -\frac{R_2}{L_2} & 0 \\ 0 & 0 & -\frac{R_3}{L_3} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} -\frac{R_1}{L_1} & 0 & 0 \\ 0 & -\frac{R_3}{L_3} & 0 \\ 0 & 0 & -\frac{R_2}{L_2} \end{bmatrix} = \begin{bmatrix} \overline{A}_1 & 0 \\ 0 & \overline{A}_2 \end{bmatrix}. \tag{5.12}$$

and

$$\overline{A}_1 = \begin{bmatrix} -\frac{R_1}{L_1} & 0 \\ 0 & -\frac{R_3}{L_3} \end{bmatrix}, \quad \overline{A}_2 = \left[-\frac{R_2}{L_2} \right]. \tag{5.13}$$

The reachable pair $(\overline{A}_1, \overline{B}_1)$ is reachable and the pair $(\overline{A}_2, \overline{B}_2 = 0)$ is unreachable.

In general case we have the following theorem [10].

Theorem 5.3. The linear electrical circuit composed of resistors, coils and voltage source is positive for any values of the resistances, inductances and source voltages if the number of coils is less or equal to the number of its linearly independent meshes and the direction of the mesh currents are consistent with the directions of the mesh source voltages.

These considerations can be extended to fractional positive electrical circuits [8].

6 Decoupling Zeros of the Positive Electrical Circuits

In a similar way as for linear systems we will define the input-decoupling zeros of the positive electrical circuits.

Definition 6.1. Let $\overline{A}_2$ be the matrix of unreachable part of the positive electrical circuit (2.1). The zeros $s_{i1}, s_{i2}, ..., s_{i\overline{n}_2}$ of the characteristic polynomial

$$\det[I_{\overline{n}_2} s - \overline{A}_2] = s^{\overline{n}_2} + \overline{a}_{\overline{n}_2 -1} s^{\overline{n}_2 -1} + ... + \overline{a}_1 s + \overline{a}_0 \tag{6.1}$$

of the matrix $\overline{A}_2$ are called the input-decoupling zeros of the positive electrical circuit (2.1).

The list of the input-decoupling zeros will be denoted by $Z_i = \{s_{i1}, s_{i2}, ..., s_{in_2}\}$.

Theorem 6.1. The state vector $x(t)$ of the positive electrical circuit is independent of the input-decoupling zeros for any input $u(t)$ and zero initial conditions. Proof is similar to the proof of Theorem 4.1.

Example 6.1. (continuation of Example 5.2) In Example 5.2 it was shown that for the unreachable pair $(\overline{A}_2, \overline{B}_2 = 0)$ the matrix $\overline{A}_2$ has the form $\overline{A}_2 = \left[-\frac{R_2}{L_2}\right]$. Therefore, by Definition 6.1 the electrical circuit shown in Figure 5.3 has one input-decoupling zero $s_{i1} = -\frac{R_2}{L_2}$. Note that the input-decoupling zero corresponds to the mesh without the source voltage (e_2 = 0).

Definition 6.2. Let $\hat{A}_2$ be the matrix of unobservable part of the positive electrical circuit (2.3). The zeros $s_{o1}, s_{o2}, ..., s_{o\hat{n}_2}$ of the characteristic polynomial

$$\det[I_{\hat{n}_2} s - \hat{A}_2] = s^{\hat{n}_2} + \hat{a}_{\hat{n}_2 -1} s^{\hat{n}_2 -1} + ... + \hat{a}_1 s + \hat{a}_0 \tag{6.2}$$

of the matrix $\hat{A}_2$ are called the output-decoupling zeros of the positive electrical circuit (2.3).

The list of the output-decoupling zeros will be denoted by $Z_o = \{s_{o1}, s_{o2}, ..., s_{o\hat{n}_2}\}$.

Theorem 6.2. The output vector $y(t)$ of the positive electrical circuit is independent of the output-decoupling zeros for any input $\overline{u}(t) = Bu(t)$ and zero initial conditions. Proof is similar to the proof of Theorem 4.1.

It is easy to show for the positive electrical circuit with $C = [0 \quad 0 \quad R_3]$ from Fig. 5.3 that the matrix $\hat{A}_2$ of the unobservable pair has the form

$$\hat{A}_2 = \begin{bmatrix} -\frac{R_1}{L_1} & 0 \\ 0 & -\frac{R_2}{L_2} \end{bmatrix}. \tag{6.3}$$

Therefore, by Definition 6.2 the positive electrical circuit shown in Fig. 5.2 has two output-decoupling zero $s_{o1} = -\frac{R_1}{L_1}$, $s_{o2} = -\frac{R_2}{L_2}$.

Definition 6.3. Zeros $s_{io}^{(1)}, s_{io}^{(2)}, ..., s_{io}^{(k)}$ which are simultaneously the input-decoupling zeros and the output-decoupling zeros of the positive electrical circuit are called the input-output decoupling zeros of the positive electrical circuit, i.e.

$$s_{io}^{(j)} \in Z_i \text{ and } s_{io}^{(j)} \in Z_o \text{ for } j = 1,2,\dots,k;\ k \le \min(\bar{n}_2, \hat{n}_2). \tag{6.4}$$

The list of input-output decoupling zeros will be denoted by $Z_{io} = \{z_{io}^{(1)}, z_{io}^{(2)}, ..., z_{io}^{(k)}\}$.

Example 6.2. Consider the positive electrical circuit shown in Fig. 5.3 with the matrices *A*, *B*, *C* given by (5.10b) and $C = [0 \quad 0 \quad R_3]$. In Example 6.1 it was shown that the positive electrical circuit has one input-decoupling zero $s_{i1} = -\frac{R_2}{L_2}$ and two output-decoupling zeros $s_{o1} = -\frac{R_1}{L_1}$, $s_{o2} = -\frac{R_2}{L_2}$. Therefore, by Definition 4.3 the positive electrical circuit has one input-output decoupling zero $s_{io}^{(1)} = -\frac{R_2}{L_2}$.

These considerations can be extended to fractional electrical circuits [8].

7 Concluding Remarks

New necessary and sufficient conditions for the reachability and observability of the positive linear electrical circuits have been established. The definitions of the input-decoupling zeros, output-decoupling zeros and input-output decoupling zeros of the positive systems and electrical circuits have been proposed. Some properties of the new decoupling zeros have been discussed. The considerations have been illustrated by numerical examples of positive electrical circuits (systems) composed of resistors, coils and voltage source. An open problem is an extension of these considerations to fractional discrete-time and continuous-time positive linear systems and fractional electrical circuits [11].

Acknowledgment. This work was supported under work S/WE/1/11.

References

1. Antsaklis, P.J., Michel, A.N.: Linear Systems. Birkhauser, Boston (2006)
2. Farina, L., Rinaldi, S.: Positive Linear Systems; Theory and Applications. J. Wiley, New York (2000)
3. Kaczorek, T.: Positive 1D and 2D systems. Springer, London (2001)
4. Kaczorek, T.: Linear Control Systems, vol. 1. J. Wiley, New York (1993)
5. Kaczorek, T.: Decomposition of the pairs (A,B) and (A,C) of the positive discrete-time linear systems. Archives of Control Sciences 20(3), 341–361 (2010)
6. Kaczorek, T.: Controllability and observability of linear electrical circuits. Electrical Review 87(9a), 248–254 (2011)
7. Kaczorek, T.: Decoupling zeros of positive discrete-time linear systems. Circuit and Systems 1, 41–48 (2010)
8. Kaczorek, T.: Positivity and reachability of fractional electrical circuits. Acta Mechanica et Automatica 5(2), 42–51 (2011)
9. Kaczorek, T.: Reachability and controllability to zero tests for standard and positive fractional discrete-time systems. Journal Europeen des Systemes Automatises, JESA 42(6-8), 770–781 (2008)
10. Kaczorek, T.: Positive electrical circuits and their reachability. Archives of Electrical Engineering 60(3), 283–301 (2011)
11. Kaczorek, T.: Selected Problems of Fractional Systems Theory. Springer, Berlin (2011)
12. Kaczorek, T.: Decoupling zeros of positive continuous-time linear systems. Bull. Pol. Acad. Sci. Tech. 61(3) (2013)
13. Kaczorek, T.: Decoupling zeros of positive electrical circuits. Archives of Electrical Engineering 62(2) (2013)
14. Kailath, T.: Linear Systems. Prentice-Hall, Englewood Cliffs (1980)
15. Kalman, R.E.: Mathematical Descriptions of Linear Systems. SIAM J. Control 1, 152–192 (1963)
16. Kalman, R.E.: On the General Theory of Control Systems. In: Proc. of the First Intern. Congress on Automatic Control, pp. 481–493, Butterworth, London (1960)
17. Rosenbrock, H.H.: State-Space and Multivariable Theory. J. Wiley, New York (1970)
18. Tokarzewski, J.: Finite zeros of positive linear discrete-time systems. Bull. Pol. Acad. Sci. Tech. 59(3), 287–292 (2011)
19. Tokarzewski, J.: Finite zeros of positive continuous-time systems. Bull. Pol. Acad. Sci. Tech. 59(3), 293–302 (2011)
20. Tokarzewski, J.: Finite Zeros in Discrete-Time Control Systems. Springer, Berlin (2006)
21. Wolovich, W.A.: Linear Multivariable Systems. Springer, New York (1974)

Information Security of Nuclear Systems

Jason T. Harris

Idaho State University, Department of Nuclear Engineering and Health Physics,
Idaho, USA
harrjaso@isu.edu

Abstract. The emphasis on information or cyber security has increased drastically over the last several years due to the increased number in attacks at the personal, company and even state level. Billions of Euros have been lost due to these attacks and the amount of funding expended to prevent them is even greater. One particular area of concern is in the protection of nuclear system information assets. These assets pertain to both information technology (IT) and instrumentation and control (I&C). Nuclear power plants (NPPs) are especially concerned as they transition from analog to digital technology. Nuclear information security garnered global attention at the recent 2012 Seoul Nuclear Security Summit and the 2013 International Conference on Nuclear Security: Enhancing Global Efforts. This paper discusses the information security domains at NPPs, the nuclear IT and I&C assets, and what these facilities are doing to protect themselves from cyber threats.

Keywords: Nuclear security, information security, nuclear power plant.

1 Introduction

The need for information security is not a new phenomenon. Encrypted messages have been used for thousands of years in areas such as warfare and politics. Cryptography, which is the study of techniques for secure communication, can be traced back to ancient Egypt and Greece. One of the earliest methods of encryption utilized a transposition cipher. The cipher contains a message in which the positions held by units of text are shifted according to a regular system. The cipher text, which has encoded or encrypted information, constitutes a permutation of the text and the order of units is changed. Some mathematical function is used on the characters' positions to encrypt, with an inverse function used to decrypt.

Since ancient times, many things have changed in regard to information security. First, both the speed at which information is processed and the throughput has increased significantly. There is also increased reliance on information and the information itself can be much more complex. There is an increased use of computerized, networked control systems and a human mediator may not filter information. Finally there are increased risks related to knowledge of information.

Information security faces new challenges in today's world due to the emergence of hacking, cyberterrorism, and cyberwarfare. One very specific area of information security where there is increasing concern is in nuclear systems. Information security

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_2, © Springer International Publishing Switzerland 2014

is an integral part of nuclear security. Nuclear information security is concerned with the protection of information assets from a wide range of threats specific to nuclear systems. The objectives of nuclear information security include: protection against loss of nuclear sensitive and/or classified information; protection against the theft of material (both physical and information); protection against terrorist actions and sabotage; protections against a combined cyber and physical attack, ensuring nuclear safety, and ensuring business continuity [3]. Consequences of an event related to nuclear information can affect the owner of the information or system, organizations and individuals responsible for secure and safe operation of the process, the government, and possibly the public.

The importance of nuclear information security was made clear at two recent and significant conferences: the Nuclear Security Summit held in Seoul, Korea in 2012 and the International Conference on Nuclear Security held in Vienna in 2013. The 2012 Nuclear Security Summit had participation from more than 53 heads of state and international organizations. The International Conference on Nuclear Security held by the International Atomic Energy Agency (IAEA) had participations from nearly 600 governmental representatives and nuclear security technical experts from all over the world. Both meetings focused on the importance of preventing theft of information, technology or expertise required to acquire and use nuclear materials for malicious purposes, or to disrupt information technology based control systems at nuclear facilities [1, 2]. The loss of nuclear information at Los Alamos National Laboratory, and recent cyber attacks on the Davis-Besse nuclear power plant and Iranian nuclear facilities are proof that the threats are real.

2 Security Concepts

Before discussing nuclear information security, it is necessary to describe basic security attributes and information assets. The three basic attributes of all security-related systems are confidentiality, integrity, and availability (CIA approach). Confidentiality means ensuring that unauthorized people, resources, or processes cannot access information. Integrity involves the protection of information from intentional or accidental unauthorized changes. Availability is assurance that information is available whenever needed. Both integrity and availability are critical for real-time control applications. For nuclear facilities, two additional security principles can be applied and are important when safety needs to be considered. These facilities need to be free from interference and be robust against undesired attacks.

Information assets are all elements of information that either shares a common usage, purpose, associated risk and/or form of storage. These assets are also defined as sets of information that are considers of value to an individual, organization, or State. Information security is the regime or program in place to ensure the protection of these assets.

2.1 Information Security Objectives

Information, or more specifically computer, security requires protecting the confidentiality, integrity and availability attributes of electronic data or computer systems and processes. By identifying and protecting these attributes in data or systems that can have an adverse impact on the safety and security functions in facilities (nuclear), the security objectives can be met [4]. Computer security, used in this context, is the same as information technology (IT) security or cyber security.

Security of computer systems, regardless of location, requires specific measures. Preventative measures are set up to protect against threats. Measures and processes must also be in place to detect threats, both single and continuous. Processes and instructions must also be in place to properly respond or react to threats or attacks. Measures must also include an evaluation of risk. Risk is the potential that a given threat will exploit the vulnerabilities of an asset or group of assets. It is measured in terms of a combination of the likelihood of an event and the severity of its consequences.

2.2 Nuclear Security Objectives

Information security at nuclear facilities is a crucial component of an overall nuclear security plan. This nuclear security plan is developed as part of a State's nuclear security regime. The International Atomic Energy Agency (IAEA) has developed nuclear security recommendations on radioactive material and associated facilities. The overall objective of a State's nuclear security regime is to protect persons, property, society, and the environment from malicious acts involving nuclear material or other radioactive material that could cause unacceptable radiological consequences. Ultimately, protection should be from both unauthorized removal (theft) of material and acts of sabotage [5].

Nuclear security relies on the identification and assessment of those assets that need to be secured. Classification of sensitive information assets defines how they should be protected. In a nuclear facility, there are three types of computer (digital) equipment assets. Information technology (IT) consists of computers, networks and databases. Industrial control systems (ICS) encompass several types of control systems used in industrial production and critical infrastructures. This includes supervisory control and data acquisition (SCADA) systems and other control system configurations such as programmable logic controllers (PLC). Instrumentation and control (I&C) systems support nuclear production processes in plants. The IAEA has determined there are three groups of threats on computer systems at nuclear facilities: information gathering attacks aimed at planning and executing further malicious attacks; attacks disabling or compromising attributes of one or several computers crucial to facility security or safety; and attacks on computers combined with a physical intrusion to target locations [4].

Implementing adequate security and countermeasures at nuclear facilities is very different than in other business sectors. The typical approach for protection of a company's assets depends on risk appetite and tolerances. Unrestricted use of this approach is not acceptable for nuclear systems. Security risks are partially determined by a State. The State develops a Design Basis Threat (DBT) that outlines security requirements and then is given to the facility. The DBT takes into account all credible, postulated, and perceived threats that various assets within a State may be vulnerable to a threat or attack. These threats may be physical (i.e. sabotage, explosion, theft) or cyber in nature. The DBT is developed by the State with input from various governmental agencies, regulators, and facilities.

3 Nuclear Information Security Domain

Operations at nuclear facilities are very unique and as such, the computer security requirements are equally unique. Nuclear operations can include several modes such as start-up, operating power, hot shutdown, cold shutdown, and refueling. Testing of systems is also done at various time intervals (i.e. daily, weekly, and monthly). The need for guidance addressing computer security at nuclear facilities is supported by the special conditions characterizing the industry. Nuclear facilities must abide by requirements set by their national regulatory bodies, which may directly or indirectly regulate computer systems or set guidance. Nuclear facilities may have to protect against additional threats, which are not commonly considered in other industries. Such threats may also be induced by the sensitive nature of the nuclear industry. Computer security requirements in nuclear facilities may differ from requirements in other concerns. Typical business operations involve only a limited range of requirements. Nuclear facilities need to take a wider base or an entirely different set of considerations into account.

Nuclear information security domains vary depending on the type of facility. This paper will not cover the multitude of nuclear facilities (enrichment facilities, fuel fabrication facilities, waste storage facilities, etc) available. Instead, the nuclear information security of a nuclear power plant (NPP) will be expounded upon. Nuclear power plants are found in over 30 countries and represent one of the most credible sources for a cyber attack. The information security domain of a typical NPP is shown in Figure 1 [6].

Information and control (I&C) assets are all digital elements that are used for safety functions or operational control, especially for measurement, actuation, and monitoring plant parameters. Examples of these assets include reactor protection systems, emergency core cooling, emergency power supply, reactor power control, fire detection, and radiation monitoring. General components of I&C systems include micro-controllers, drivers, motors, valves, servers, and gateways. Most information security processes are focused on IT environments. However, great changes to digital based technologies on I&C systems are occurring in nuclear facilities. Previously control systems associated with nuclear power plants existed as isolated “islands”, but that has changed. In a NPP control room, the operator workstation will be connected by a number of operations, maintenance, data acquisition, and process handling data highways. Control and protection related systems are independent of these data highways. Vital plant control systems will utilize redundant communications. Protection systems will utilize redundant channels.

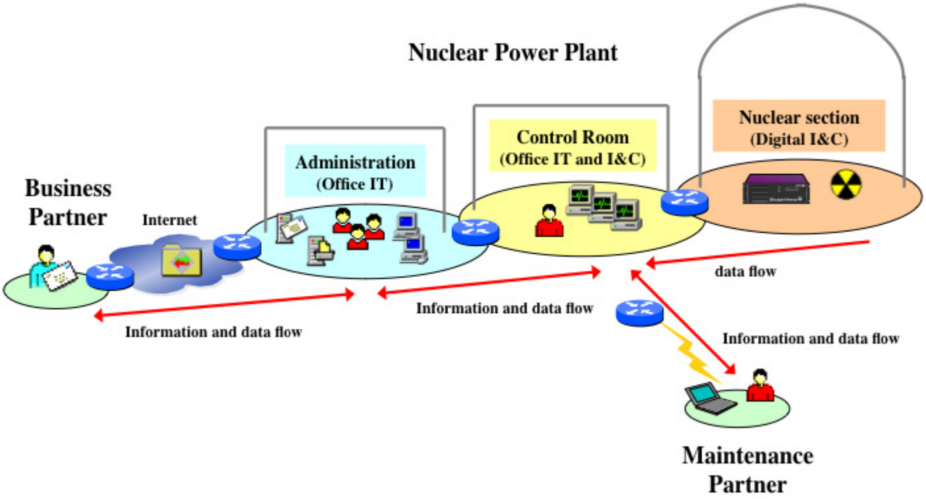

Nuclear Power Plant
Business Partner
Internet
Administration (Office IT)
Control Room (Office IT and I&C)
Nuclear section (Digital I&C)
data flow
Information and data flow
Information and data flow
Information and data flow
Maintenance Partner

Fig. 1. Information security domains at a nuclear power plant (Source: INSEN, p. 7)

The importance of I&C asset security cannot be overstated. But information security at a NPP is equally important. At the core of information security is the development of an information security management system (ISMS). The ISMS means: understanding an organization’s information security requirements and the need to establish policy and objectives for information security; implementing and operating controls to manage an organization's information security risks in the context of the organization’s overall business risks; monitoring and reviewing the performance and effectiveness of the ISMS; and continual improvement based on objective measurement. At a NPP, the specific drivers to the ISMS process include generally increased IT security requirements (critical infrastructure), legal requirements, directives and guidelines, co-determination by authorities, and consideration of insider threat. Subject areas focus on management, business processes, the plant, operations, information systems, access control, and the workforce.

4 Protecting Nuclear Information Assets

Protecting information assets requires the integration of several approaches. First, a holistic approach must be taken. Implementation of this approach requires identifying and classifying all information assets, assessing threats, vulnerabilities and risk, and applying graded protective measures according to the asset’s life cycle. All disciplines of security at the facility must interact and complement each other to establish a facility’s security posture. Next, human factors must be considered. Key to this is the development of a strong security culture. Security culture is a set of beliefs, attitudes, behaviors, and management systems that lead to a more effective nuclear security program. This includes information /computer security culture. As humans are typically the weakest link when it comes to security, ongoing training and strong management is very important. Heavy security in nuclear facilities must also be balanced with acceptance by the workforce.

There are a number of concepts that are utilized to protect nuclear information. Access control, authorization, and need-to-know concepts are used extensively in a variety of business environments. Nuclear facilities also incorporate a graded approach to security as well as defense-in-depth. The graded approach uses security measures that are adequate to the protection level needed. As the criticality of the systems increases, so do the level of security/strength of measures. The defense-in-depth principle arranges information assets in a way that low sensitive assets are easier to access than high sensitive assets (i.e. zone model with subordinate barriers). In a NPP, there will be internet and network zone borders. Technical concepts focus on the protection of the information assets itself and of the infrastructure the information assets use. Technical measures impact organizational issues like IT processes. Examples of technical concepts include network security, authentication and crypotography, intrusion detection, and network management. Finally, organizational concepts focus on the processes dealing with the information assets. Security must be implemented in the processes that are usually driven by the IT department. Subject areas include purchasing, software development, problem management, incident management, and help desk.

Nuclear facilities have a number of special considerations that need to be addressed when it comes to information security. First, facility lifetime phases and modes of operation must be evaluated. Access of information may vary drastically at these different times. Also there will be security requirement differences between IT and I&C systems. The potential consequences that come with the demand for additional connectivity must be carefully considered. Software updates, secure design, and specifications for computer systems throughput the facility must be evaluated. Finally, third party or vendor access control procedures must be developed, evaluated, and updated on a continuous basis.

5 Conclusions

As the number of cyber and information attacks increases, facilities must be increasingly vigilant of protecting their assets. This is especially true for nuclear facilities where compromised security can lead to degradation of safety systems, which in turn can lead to detrimental consequences to the facility, humans, and the environment. It is of utmost importance that information security professionals are trained properly and can stay ahead of the threat. Educational initiatives, much like what has been developed by the IAEA, are crucial for developing a workforce that can keep nuclear facilities safe and secure from malicious acts.

References

1. Nuclear Security Summit: Seoul Communiqué. Nuclear Security Summit. IAEA, Vienna (March 27, 2012)
2. International Atomic Energy Agency: Proceedings of the International Conference on Nuclear Security: Enhancing Global Efforts. IAEA, Vienna (2013)

3. International Atomic Energy Agency: Introduction to Information and Computer Security. IAEA Introduction to Nuclear Security Module 14. IAEA, Vienna (2012)
4. International Atomic Energy Agency: Computer Security at Nuclear Facilities. IAEA Nuclear Security Series No. 17. IAEA, Vienna (2011)
5. International Atomic Energy Agency: Nuclear Security Recommendations on Radioactive Material and Associated Facilities. IAEA Nuclear Security Series No. 14. IAEA, Vienna (2011)
6. International Atomic Energy Agency: NS 1 Introduction to Nuclear Security: NS 1.11 Information Security. INSEN Educational Material, pp. 1–60. IAEA, Vienna (2012)
7. International Atomic Energy Agency: Core Knowledge on Instrumentation and Control Systems in Nuclear Power Plants. IAEA Nuclear Energy Series No. NP-T-3.12. IAEA, Vienna (2011)

Distributed Reconfigurable Predictive Control of a Water Delivery Canal

João M. Lemos[1], José M. Igreja[2], and Inês Sampaio[1]

[1] INESC-ID/IST, Technical University of Lisbon, Lisboa, Portugal
jlml@inesc-id.pt
[2] INESC-ID/ISEL, Lisboa, Portugal
jigreja@deea.isel.ipl.pt

Abstract. This paper addresses the problem of reconfigurable distributed model predictive control (MPC) of water delivery canals. It is shown how a distributed MPC algorithm can be equipped with complementary features so as to reconfigure its structure in order to render it tolerant to actuator faults. The structure proposed includes a fault detection algorithm that triggers switching between different controllers designed to match the fault or no-fault situation. To ensure stability, a dwell-time switching logic is used. Experimental results are provided.

Keywords: Fault tolerant control, reconfigurable control, distributed control, predictive control, water delivery canal.

1 Introduction

Water delivery open canals used for irrigation [1] are large structures whose complexity, together with increasing requirements on reliability and quality of service provides a strong motivation to consider fault tolerant control methods [2]. In order to achieve fault tolerant features, the idea consists in exploring the redundancy in instaled sensors and actuators to reconfigure the control system such as to allow the plant operation to continue, perhaps with some graceful degradation, when a sensor or actuator fails.

The concept of fault tolerant control (FTC) has been the subject of intense research in the last twenty years [3–5], in particular in what concerns reconfigurable fault tolerant control systems [7]. This activity yielded a rich bibliography that, of course, cannot be covered here and that comprises aspects such as fault detection and isolation and fault tolerant control design. In relation to distributed control, an important concept is "integrity", namely the capacity of the system to continue in operation when some part of it fails [6]. Other type of approach models the failures as disturbances that are estimated and compensated by the controller [8]. In what concerns water delivery canal systems topics found in the literature include control loop monitoring [13] and reconfiguration to mitigate fault effects [12]. Reconfiguring the controller in face of a plant fault falls in the realm of hybrid systems and raises issues related to stability that must be taken into account [10].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_3, © Springer International Publishing Switzerland 2014

The contribution of this paper consists of the application of MPC based distributed fault tolerant control to a water delivery canal in the presence of actuator faults. An algorithm based on controller reconfiguration with a dwell time logic is presented, together with experimental results.

The paper is organized as follows: After the introduction in which the work is motivated, a short literature review is made and the main contributions are presented, the canal is described in section 2, including a static nonlinearity compensation of the gate model. Distributed MPC control is described in section 3, whereas actuator fault tolerant control is dwelt with in section 4. Experimental results are presented in section 5. Finally, section 6 draws conclusions.

2 The Canal System

2.1 Canal Description

The experimental work reported hereafter was performed at the large scale pilot canal of *Núcleo de Hidráulica e Controlo de Canais* (Universidade de Évora, Portugal), described in [11]. The canal has four pools with a length of 35m, separated by three undershoot gates, with the last pool ended by an overshoot gate. In this work, only the first three gates are used. The maximum nominal design flow is $0.09\,\mathrm{m}^3\mathrm{s}^{-1}$. There are water off-takes downstream from each branch made of orifices in the canal walls, that are used to generate disturbances corresponding to water usage.

Water level sensors are installed downstream of each pool. The water level sensors allow to measure values between 0 mm and 900 mm, a value that corresponds to the canal bank. For pool number i, $i = 1, \ldots, 4$, the downstream level is denoted y_i and the opening of gate i is denoted u_i. The nomenclature is such that pool number i ends with gate number i. Each of the actual gate positions $u_{r,i}$, $i = 1, 2, 3$ is manipulated by a command signal u_i.

2.2 Nonlinearity Compensation

Following [2], in order to compensate for a nonlinearity, instead of using as manipulated variable the gate positions $u_{r,i}$, the corresponding water flows q_i crossing the gates are used. These are related by

$$q_i = C_{ds} W u_{r,i} \sqrt{2g(h_{upstr,i} - h_{downstr,i})}, \tag{1}$$

where C_{ds} is the discharge coefficient, W is the gate width, $g = 9,8\mathrm{m/s}$ is the gravity acceleration, $h_{upstr,i}$ is the water level immediately upstream of the gate and $h_{downstr,i}$ is the water level immediately downstream of the gate. This approach corresponds to representing the canal by a Hammerstein model and to compensating the input nonlinearity using its inverse. The linear controller computes the flow crossing the gates, that is considered to be a virtual command variable v_i and the corresponding gate position is then computed using (1). The discharge coefficient is not estimated separately, but instead is considered to be incorporated in the static gain of the linear plant model.

2.3 Canal Model

In order to design the controllers, the dynamics of the canal has been approximated by a finite dimension linear state-space model written as

$$x(k+1) = Ax(k) + Bv(k), \tag{2}$$

$$y(k) = Cx(k) \tag{3}$$

where $k \in \mathbb{N}$ denotes discrete time, $x \in \mathbb{R}^n$ is the full canal state, $y \in \mathbb{R}^p$ is the output made of the downstream pool levels, with $p = 3$ the number of outputs, $v \in \mathbb{R}^p$ is the manipulated variable and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$ and $C \in \mathbb{R}^{p\times n}$ are matrices. Assuming operation around a constant equilibrium point, these matrices are identified by constraining the model to have the following structure

$$A = \begin{bmatrix} A_{11} & \underline{0} & \underline{0} \\ \underline{0} & A_{22} & \underline{0} \\ \underline{0} & \underline{0} & A_{33} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} & \underline{0} \\ B_{21} & B_{22} & B_{23} \\ \underline{0} & B_{32} & B_{33} \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & \underline{0} & \underline{0} \\ \underline{0} & C_2 & \underline{0} \\ \underline{0} & \underline{0} & C_3 \end{bmatrix}. \tag{4}$$

These matrices have dimensions that match the state x_i associated to each pool such that $x = [x_1' x_2' x_3']'$. This structure is imposed to reflect the decomposition of the canal model in subsystems, each associated to a different pool. Furthermore, it is assumed that each pool interacts directly only with its neighbors, and only through the input.

3 Distributed Siorhc

3.1 A Strategy for Distributed Control

Let the canal be decomposed in a number of local linear time invariant subsystems Σ_i, $i = 1, \ldots, N_\Sigma$. Each of these subsystems have a manipulated input u_i and a measured output y_i. In addition, Σ_i interacts with its neighbors, Σ_{i-1} and Σ_{i+1}. It is assumed that this interaction takes only part through the manipulated variables and that the cross-coupling of states can be neglected.

As shown in figure 1, a local controller $\mathscr{C}_i$ is associated to each subsystem Σ_i. At the beginning of each sampling interval, this local controller computes the value of the manipulated variable using knowledge of y_i but also by performing a negotiation with the adjacent controllers $\mathscr{C}_{i-1}$ and $\mathscr{C}_{i+1}$. This negotiation takes place in a recursive way along the following steps:

1. Let l be the step index and $\Delta U_i(k, l)$ be the increment of the manipulated variable of local controller i at sampling time k and after performing l steps.
2. Set the counter $l = 1$.
3. Assume that each local controller i, $i = 1, 2, 3$, at time k and after performing l steps knows $\Delta U_{i-1}(k, l-1)$ and $\Delta U_{i+1}(k, l-1)$, *i. e.*, each local controller knows the previous iteration of the neighbor controllers. Update the control increment of each local controller by

$$\Delta U_i(k, l) = \mathscr{F}(\Delta U_{i-1}(k, l-1), \Delta U_{i+1}(k, l-1)) \tag{5}$$

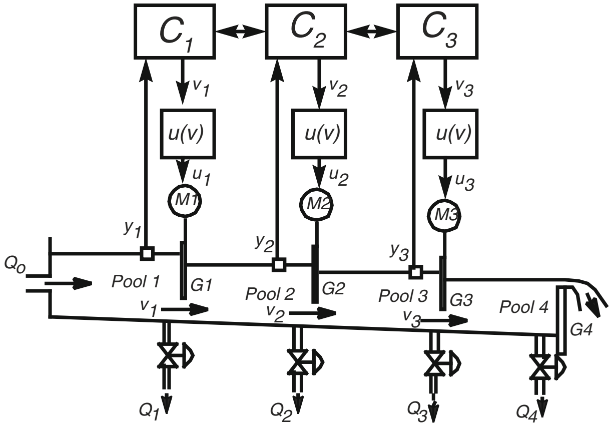


Fig. 1. Distributed controller in normal (no fault) operation

where $\mathscr{F}$ denotes the optimization procedure used, that varies from algorithm to algorithm.

4. If convergence is reached, stop. Otherwise, set $l \rightarrow l+1$ and go to step 2.

In the next sub-section, a distributed version of a model predictive controller with stability constraints, named SIORHC, is obtained using this procedure.

3.2 Distributed Siorhc

Consider the system described by the linear state model (2), augmented with an integrator. For this system, the predicted outputs at $k+j$ given observations up to time k are given by:

$$\begin{aligned}\hat{y}(k+j) &= \sum_{i=0}^{j-1} C\ A^{j-i-1} B \varDelta u(k+i) + \hat{y}_0(k+j) \\ \hat{y}_0(k+j) &= C\ A_j\ \hat{x}(k)\end{aligned} \tag{6}$$

where $\hat{y}_0$ is the output predict value without control moves (the system free response) and $\hat{x}$ denote either the state or its estimate obtained with a suitable observer. For $j = 1...N,\ N+1, ...,\ N+P$ (6) the predictors can be written in a compact way as

$$\begin{aligned}\hat{Y}_N &= G_N \varDelta U + \hat{Y}_{0N} \\ \hat{Y}_P &= G_P \varDelta U + \hat{Y}_{0P}\end{aligned} \tag{7}$$

with

$$\begin{aligned}\hat{Y}_N &= \left[y_{k+1} \cdots y_{k+N}\right]^T \\ \hat{Y}_P &= \left[y_{k+N+1} \cdots y_{k+N+P}\right]^T\end{aligned} \tag{8}$$

In order to develop a distributed controller version, let the system be decomposed in a number of serially connected subsystems Σ_i, $i = 1, \ldots, N_\Sigma$. For the sake of clarity consider the case $N_\Sigma = 3$. Equation (7) is then approximated considering only interactions between neighboring serially connected systems:

$$\begin{aligned}
\hat{Y}_{1N} &= \hat{Y}_{10N} + G_{11N}\Delta U_1 + G_{12N}\Delta U_2\\
\hat{Y}_{1P} &= \hat{Y}_{10P} + G_{11P}\Delta U_1 + G_{12P}\Delta U_2\\
\hat{Y}_{2N} &= \hat{Y}_{20N} + G_{21N}\Delta U_1 + G_{22N}\Delta U_2 + G_{23N}\Delta U_3\\
\hat{Y}_{2P} &= \hat{Y}_{20P} + G_{21P}\Delta U_1 + G_{22P}\Delta U_2 + G_{23P}\Delta U_3\\
\hat{Y}_{3N} &= \hat{Y}_{30N} + G_{32N}\Delta U_2 + G_{33N}\Delta U_3\\
\hat{Y}_{3P} &= \hat{Y}_{30P} + G_{32P}\Delta U_2 + G_{33P}\Delta U_3
\end{aligned}$$

Associate the following local cost functional to Σ_1:

$$\overline{J_1} = \sum_{i=1}^{N} e_{1,\,k+i}^T Q_1 e_{1,\,k+i} + \sum_{i=1}^{N} e_{2,\,k+i}^T Q_2 e_{2,\,k+i} + \sum_{i=0}^{N-1} \Delta u_{1,\,k+i}^T R_1 \Delta u_{1,\,k+i} \tag{9}$$

with zero terminal horizon constraint given by:

$$\left[y_{1,\,k+N+1} \ \cdots \ y_{1,\,k+N+P}\right]^T = \left[r_{1,\,k+N+1} \ \cdots \ r_{1,\,k+N+P}\right]^T \tag{10}$$

where $e_{(.),k} = r_{(.),k} - y_{(.),k}$, is the tracking error in relation to the reference sequence, $r_{(.),k}$, and $Q_{(.)} \geq 0$ and $R_{(.)} > 0$ are weighting matrices.

In an equivalent way, the minimization of $\overline{J_1}$ with respect to ΔU_1 may be written as:

$$\min_{\Delta U_1} \overline{J_1} = \|Y_{1RN} - \hat{Y}_{1N}\|_{Q_1}^2 + \|Y_{2RN} - \hat{Y}_{2N}\|_{Q_2}^2 + \|\Delta U_1\|_{R_1}^2 \tag{11}$$

$$s.t. \quad \hat{Y}_{1P} = Y_{1RP}$$

The stated QP optimization problem with constraints can now be solved by finding the vector ΔU_1 that minimizes the Lagrangian:

$$\begin{aligned}
\mathscr{L}_1 := {} & \|E_1 - G_{11N}\Delta U_1 - G_{12N}\Delta U_2\|_{Q1}^2 + \\
& + \|E_2 - G_{21N}\Delta U_1 - G_{22N}\Delta U_2 - G_{23N}\Delta U_3\|_{Q2}^2 + \\
& + \|\Delta U_1\|_{R1}^2 + \|\Delta U_2\|_{R2}^2 + \\
& + \left[F_1 + G_{11P}\Delta U_1 + G_{12P}\Delta U_2\right]^T \lambda_1
\end{aligned} \tag{12}$$

where $E_j = Y_{jRN} - \hat{Y}_{j0N}$, $F_j = Y_{jRP} - \hat{Y}_{j0P}$ and λ_1 is a column vector of Lagrange multipliers. Solving (12) yields

$$M_1 \Delta U_1 = \left(I - G_{11P}^T W_1 G_{11P} M_1^{-1}\right)\left(G_{11N}^T Q_1 E_1 + G_{21N}^T Q_2 E_2\right) + W_1 F_1 \tag{13}$$

with

$$W_1 = \left(G_{11P} M_1^{-1} G_{11P}^T\right)^{-1} \tag{14}$$

$$M_1 = G_{11N}^T Q_1 G_{11N} + G_{21N}^T Q_2 G_{11N} + R_1 \tag{15}$$

Using analogous procedures, another two equations are obtained for the controllers associated with Σ_2 and Σ_3 by minimizing the local functionals:

$$\min_{\Delta U_2} \overline{J_2} = \|Y_{1RN} - \hat{Y}_{1N}\|^2_{Q_1} + \|Y_{2RN} - \hat{Y}_{2N}\|^2_{Q_2} + \|Y_{3RN} - \hat{Y}_{3N}\|^2_{Q_3} + \|\Delta U_2\|^2_{R_2} \quad (16)$$

$$s.t. \quad \hat{Y}_{2P} = Y_{2RP}$$

and

$$\min_{\Delta U_3} \overline{J_3} = \|Y_{2RN} - \hat{Y}_{2N}\|^2_{Q_2} + \|Y_{3RN} - \hat{Y}_{3N}\|^2_{Q_3} + \|\Delta U_3\|^2_{R_3} \quad (17)$$

$$s.t. \hat{Y}_{3P} = Y_{3RP}$$

yielding

$$\begin{aligned} M_2 \Delta U_2 = \left(I - G^T_{22P} W_2 G_{22P} M_2^{-1}\right) \left(G^T_{12N} Q_1 E_1 + \right. \\ \left. + G^T_{22N} Q_2 E_2 + G^T_{32N} Q_3 E_3\right) + G^T_{22P} W_2 F_2 \end{aligned} \quad (18)$$

and

$$\begin{aligned} M_3 \Delta U_3 = \left(I - G^T_{33P} W_3 G_{33P} M_3^{-1}\right) \left(G^T_{23N} Q_2 E_2 + \right. \\ \left. + G^T_{33N} Q_3 E_3\right) + G^T_{33P} W_3 F_3 \end{aligned} \quad (19)$$

The distributed SIORHC solution for the serially connected sub-systems can be obtained, using the procedure in subsection 3.1, from the matrix algebraic equations system:

$$\Phi \Delta U = \Psi \quad (20)$$

where the Φ matrix building blocks are:

$$\begin{aligned}
\Phi_{11} &= M_1 \\
\Phi_{12} &= S_1 \left(G^T_{11N} Q_1 G_{12N} + G^T_{21N} Q_2 G_{22N}\right) + G^T_{11P} W_1 G_{12P} \\
\Phi_{13} &= S_1 \left(G^T_{21N} Q_2 G_{23N}\right) \\
\Phi_{21} &= S_2 \left(G^T_{12N} Q_1 G_{11N} + G^T_{22N} Q_2 G_{21N}\right) + G^T_{22P} W_2 G_{21P} \\
\Phi_{22} &= M_2 \\
\Phi_{23} &= S_2 \left(G^T_{22N} Q_2 G_{23N} + G^T_{32N} Q_3 G_{33N}\right) + G^T_{22P} W_2 G_{23P} \\
\Phi_{31} &= S_3 \left(G^T_{23N} Q_2 G_{21N}\right) \\
\Phi_{32} &= S_3 \left(G^T_{23N} Q_2 G_{22N} + G^T_{33N} Q_3 G_{32N}\right) + G^T_{33P} W_3 G_{32P} \\
\Phi_{33} &= M_3
\end{aligned} \quad (21)$$

with $S_i = I - G^T_{iiP} W_i G_{iiP} M_i^{-1}$, $i = 1, 2, 3$, the entries of Ψ

$$\begin{aligned}
\Psi_1 &= S_1 \left(G^T_{11N} Q_1 A_1 + G^T_{21N} Q_2 A_2\right) + G^T_{11P} W_1 B_1 \\
\Psi_2 &= S_2 \left(G^T_{12N} Q_1 A_1 + G^T_{22N} Q_2 A_2 + G^T_{32N} Q_3 A_3\right) + G^T_{22P} W_2 B_2 \\
\Psi_3 &= S_3 \left(G^T_{23N} Q_2 A_2 + G^T_{33N} Q_3 A_3\right) + G^T_{33P} W_3 B_3
\end{aligned} \quad (22)$$

and

$$\Delta U = \left[\Delta U_1 \; \Delta U_2 \; \Delta U_3\right] \quad (23)$$

To apply the iterative procedure described in subsection 3.1, write (20) as

$$\Phi_d \Delta U(k,\, l+1) + \Phi_{nd} \Delta U(k,\, l) = \Psi \tag{24}$$

where

$$\Phi_d = \begin{bmatrix} \Phi_{11} & 0 & 0 \\ 0 & \Phi_{22} & 0 \\ 0 & 0 & \Phi_{11} \end{bmatrix} \quad \Phi_{nd} = \begin{bmatrix} 0 & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & 0 & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & 0 \end{bmatrix}. \tag{25}$$

The algorithm will converge provided that the spectral radius

$$\lambda_{max} := \max\ eig\left(\Phi_d^{-1} \Phi_{nd}\right) \tag{26}$$

verifies

$$|\lambda_{max}| < 1 \tag{27}$$

4 Fault Tolerant Control

4.1 Controller Reconfiguration

Figure 2 shows a discrete state diagram that explains how controller reconfiguration is performed when an actuator fault occurs in the water channel considered in this paper. For simplicity, only the occurrence of faults in gate 2 are considered. State $\mathcal{S}_1$ corresponds to the situation in which all gates are working normally with a controller $\mathcal{C}_N$ that matches this situation. When a fault occurs, the system state switches to $\mathcal{S}_2$, in which gate 2 is faulty (blocked) but the controller used is still the one designed for the no fault situation.

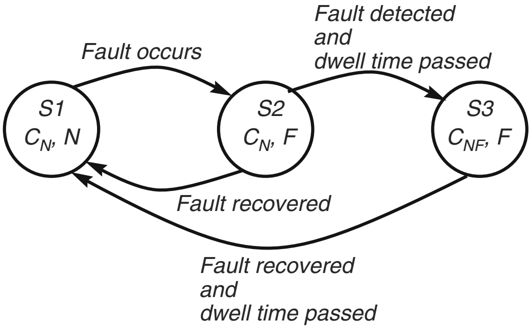

Fault occurs
Fault detected and dwell time passed
S1
C_N, N
S2
C_N, F
S3
C_NF, F
Fault recovered
Fault recovered and dwell time passed

Fig. 2. Discrete states in controller reconfiguration

In the presence of a fault, the matrices of the state-space model (3) have the structure

$$A = \begin{bmatrix} A_{11}^F & \underline{0} \\ \underline{0} & A_{33}^F \end{bmatrix}, \quad B = \begin{bmatrix} B_{11}^F & B_{13}^F \\ B_{31}^F & B_{33}^F \end{bmatrix}, \quad C = \begin{bmatrix} C_1^F & \underline{0} \\ \underline{0} & C_3^F \end{bmatrix}. \tag{28}$$

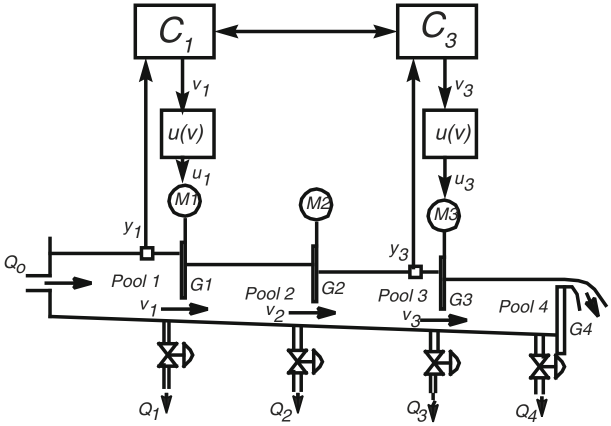


Fig. 3. Distributed controller structure after a fault is detected

The superscript F enhances the fact that the matrix blocks are estimated assuming that a fault has occurred and that they are different from the ones in (4). Figure 3 shows the controller to apply under a faulty situation. Controller reconfiguration implies a reconfiguration of the communication network as well.

When the fault is detected, the state switches to $\mathcal{S}_3$, in which a controller $\mathcal{C}_F$ designed for the faulty situation is connected to the canal. When the fault is recovered (gate 2 returns to normal operation), the state returns to $\mathcal{S}_1$. A dwell time condition is imposed to avoid instability that might arise due to fast switching [14]. This means that, once a controller is applied to the plant, it will remain so for at least a minimum time period (called dwell time). Furthermore, an integrator in series with the plant ensures bumpless transfer between controllers.

When distributed control is used, the controller designed for normal operation (shown in figure 1), $\mathcal{C}_N$, consists of 3 SISO SIORHC controllers $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$, each regulating a pool and such that each individual controller negotiates the control variable with its neighbors. This means that, in states $\mathcal{S}_1$ and $\mathcal{S}_2$, $\mathcal{C}_1$ negotiates with $\mathcal{C}_2$, $\mathcal{C}_2$ negotiates with $\mathcal{C}_1$ and with $\mathcal{C}_3$ and $\mathcal{C}_3$ negotiates with $\mathcal{C}_2$. The controller for the faulty condition (shown in figure 3) is made just of two SISO controllers that control pools 1 and 3 and negotiate with each other.

4.2 Fault Detection

For actuator faults, the fault detection algorithm operates as follows. For each gate i, $i = 1, 2, 3$, define the error $\tilde{u}_i$ between the command of the gate position u_i and the actual gate position $u_{r,i}$,

$$\tilde{u}_i(k) = u_i(k) - u_{r,i}(k) \tag{29}$$

A performance index Π is computed from this error by

$$\Pi(k) = \gamma \Pi(k-1) + (1-\gamma)|\tilde{u}(k)|. \qquad (30)$$

If $\Pi(k) \geq \Pi_{max}$, where Π_{max} is a given threshold, then it is decided that a fault has occurred.

5 Experimental Results

Figure 4 show experimental results with D-SIORHC. At the time instant marked by a red vertical line, a fault that forces gate 2 to become stuck occurs. Shortly after, at the instant marked by the yellow vertical line, this fault is detected, and the controller is reconfigured as explained. From this moment on, there is no warranty on the value of the level J_2, but J_1 and J_3 continue to be controlled.

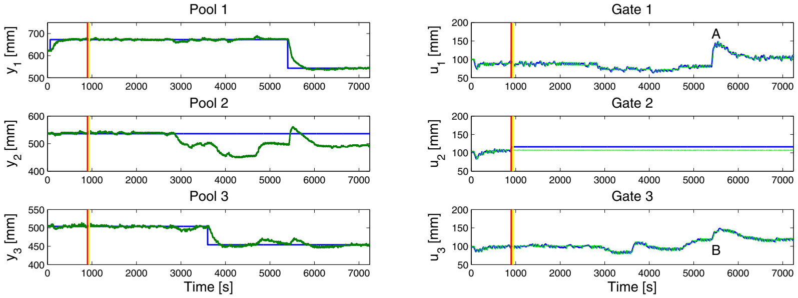


Fig. 4. D-SIORHC of the water delivery canal. Reconfiguration after a fault in gate 2.

This experiment also includes a test with respect to disturbance rejection. At time $t = 2800$ s, a disturbance is created by opening the side take of pool 1, thereby extracting some water. The controllers react by closing the gates in order to compensate for this loss of incoming flow.

6 Conclusions

A reconfigurable MPC controller with stability terminal constraints for a water delivery canal has been developed and demonstrated experimentally. Fault tolerance is embedded in the reconfiguration of a network of local controllers, that change their pattern of negotiation with neighbors in order to reach a consensus. Stability of this switched controller network is ensured by forcing a dwell time switching logic. It is possible to provide lower bounds on the dwell time that ensure stability of the overall system.

The experimental results presented illustrate that the system is able to tackle both the problems of disturbance rejection and reference tracking.

One difficulty stems from the fact that the PLCs that control gate motors impose a quantization effect. Therefore, the gate position changes only if the order for the new position differs from the actual position by at least 5 mm. This quantization effect imposes a limit on the performance of the overall system.

Acknowledgment. This work was supported by “FCT–Fundação para a Ciência e a Tecnologia”, within projects *ORCHESTRA – Distributed Optimization and Control of Large Scale Water Delivery Systems* , contract PTDC/EMS-CRO/2042/2012, and PEst-OE/EEI/LA0021/2013.

References

1. Litrico, X., Fromion, V.: Modeling and control of hydrosystems. Springer (2009)
2. Cantoni, M., Weyer, E., Li, Y., Ooi, S.K., Mareels, I., Ryan, M.: Control of large-scale irrigation networks. Proc. IEEE 95(1), 75–91 (2007)
3. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: Diagnosis and Fault Tolerant Control, 2nd edn. Springer (2006)
4. Åstrom, K.J., Albertos, P., Blanke, M., Isidori, A., Schaufelberger, W., Sanz, R. (eds.): Control of complex systems. Springer (2001)
5. Blanke, M., Staroswiecki, M., Wu, N.E.: Concepts and methods in fault-tolerant control. In: Proc. 2001 Am. Control Conf., pp. 2606–2620 (2011)
6. Campo, P.J., Morari, M.: Achievable closed-loop properties of systems under decentralized control: Conditions involving the steady-state gain. IEEE Trans. Autom. Control 39(5), 932–943 (1994)
7. Zhang, Y., Jiang, J.: Bibliographical review on reconfigurable fault-tolerant control systems. Annual Reviews in Control 32, 229–252 (2008)
8. Zhao, Q., Jiang, J.: Reliable state feedback control design against actuator failures. Automatica 34(10), 1267–1272 (1998)
9. Weyer, E., Bastin, G.: Leak detection in open water channel. In: Proc. 17th IFAC World Congress, Seoul, Korea, pp. 7913–7918 (2008)
10. Koutsoukos, X.D., Antsaklis, P.J., Stiver, J.A., Lemmon, M.D.: Supervisory control of hybrid systems. Proc. of the IEEE 88(7), 1026–1049 (2000)
11. Lemos, J.M., Machado, F., Nogueira, N., Rato, L., Rijo, M.: Adaptive and non-adaptive model predictive control of an irrigation channel. Networks and Heterogeneous Media 4(2), 303–324 (2009)
12. Choy, S., Weyer, E.: Reconfiguration scheme to mitigate faults in automated irrigation channels. In: Proc. 44th IEEE Conf. Decision and Control, Sevilla, Spain, pp. 1875–1880 (2005)
13. Zhang, P.Z., Weyer, E.: A reference model approach to performance monitoring of control loops with applications to irrigation channels. Int. J. Adaptive Control Sig. Proc. 19(10), 797–818 (2005)
14. Liberson, D., Morse, A.S.: Basic problems in stability and design of switched systems. IEEE Control Systems 19(5), 59–70 (1999)

Estimation for Target Tracking Using a Control Theoretic Approach – Part I

Stephen C. Stubberud[1,*] and Arthur M. Teranishi[2]

[1] Oakridge Technology, Del Mar, CA, United States of America
scstubberud@ieee.org
[2] Asymmetric Associates, Long Beach, CA, United States of America
art_teranishi@cox.net

Abstract. In target tracking, the estimation problem is generated using the kinematic model of the target. The standard model is the straight-line motion model. Many variants to incorporate target maneuvers have been tried including interacting multiple motion models, adaptive Kalman filters, and neural extended Kalman filters. Each has performed well in a variety of situations. One problem with all of these approaches is that, without observability, the techniques often fail. In this paper, the first step in the development of a control-loop approach to the target-tracking problem is presented. In this effort, the use of a control law in conjunction with the estimation problem is examined. This approach is considered as the springboard for incorporating intelligence into the tracking problem without using ad hoc techniques that deviate from the underpinnings of the Kalman filter.

1 Introduction

One of the primary algorithms in a target tracking system is that of a state estimation routine [2]. When measurements are assigned to a specific target, they are incorporated into the target track via the estimation algorithm. The standard estimation algorithm is that of the Kalman filter or one of its many variants such as the extended Kalman filter (EKF) [5,9]. When the target is not maneuvering and fully observable measurements of the targets are provided, the Kalman filter provides a quality estimate of the target's kinematics, position and velocity. When the target maneuvers techniques such as the interactive multiple model (IMM) approach [3,10] and the neural extended Kalman filter [11, 12] have been employed to improve the performance of the tracking system. They improve the estimate and reduce delays in the measurements keeping up with the target.

When more complex issues occur, e.g., constraints and sensors that lack observability, often techniques are developed that are ad hoc in nature. These methods are developed in such a way that the authors claim that the technique is still an EKF even though it violates the basic tenets of the Kalman filter [14]. The estimation algorithm is dependent on the kinematic model of the target being tracked. The complexity of the model increases the implementation cost of the estimator. While a number of estimation algorithm have been developed such as the constrained estimator [13] and particle filters [6,8], they are still based on estimation theory.

[*] Corresponding author.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_4, © Springer International Publishing Switzerland 2014

Instead of modifying the estimator to solve the problem, in this paper, the use of a feedback control approach is used to improve the state estimate using the EKF algorithm. Instead of estimating the velocity vector of the target, a control approach is used to modify the velocity so as to better estimate the position of the target. This is a similar approach to that of satellite guidance [7]. The control approach is a better method to handle unobservability and constraints in that more complexities can be incorporated using control theory.

This paper is the first of three that investigates the use of control laws to improve the estimation of a target. In this paper, the estimation of the target's position for a straight-line and a maneuvering target is compared using a range-bearing sensor that uses the standard EKF approach to that of a control-based implementation of the estimator. This effort investigates if the technique can provide similar results to that of the EKF that uses a full state estimate when fully observable measurements are available. With the demonstration of the baseline system, the development of a system with unobservable measurements, i.e., an angle-only sensor, can be developed. The goal is to develop a system that can work well in standard tracking applications but then be modified to overcome the problems when unobservability occurs. The third step derives a more complex control law which can incorporate a number of rules and considerations.

In Section 2, the original EKF approach is derived and then modified for the use of the control law. Section 3 describes the exemplar cases. The results are presented and discussed in Section 4. The overview of the next steps in this research outlined in Section 5.

2 Extended Kalman Filter Designs

The standard EKF for tracking is based on a kinematic model of the target dynamics. As seen in Figure 1, there exist only internal feedback to the system. The prediction component is defined as

$$\mathbf{x}_{k+1|k} = \mathbf{F}\mathbf{x}_{k|k} \tag{1}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}\mathbf{P}_{k|k}\mathbf{F}^T + \mathbf{Q}_k \tag{2}$$

where the kinematic model, the state-coupling matrix **F**, is defined as

$$\mathbf{F} = \begin{bmatrix} 1 & dt & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & dt & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & dt \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

and the state vector **x** is

$$\mathbf{x}^T = \begin{bmatrix} x & \dot{x} & y & \dot{y} & z & \dot{z} \end{bmatrix} \tag{4}$$

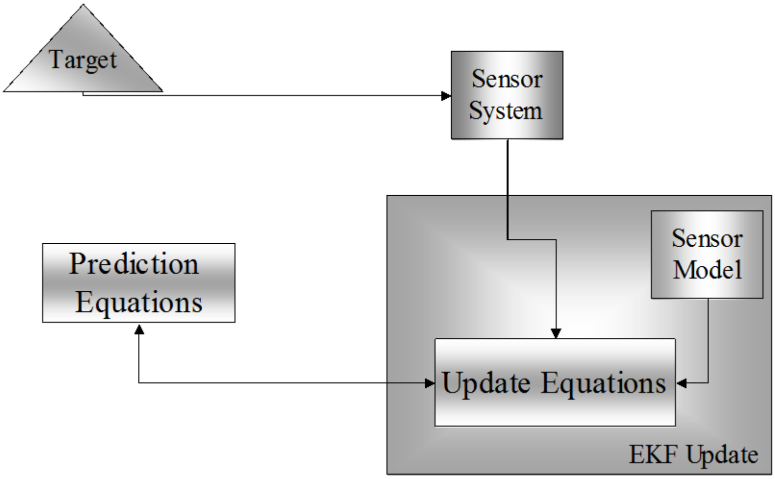


Fig. 1. The EKF is standard observer feedback loop with covariance information

This is a straight-line motion model. It is usually the best estimate of general target. Since the target is noncooperative in nature, any maneuver must be estimated through the process noise **Q.** There is no external input that can drive the system. The process noise for a tracker is often given as integrated white noise

$$\mathbf{Q} = q^2 \begin{bmatrix} \frac{dt^3}{3} & \frac{dt^2}{2} & 0 & 0 & 0 & 0 \\ \frac{dt^2}{2} & dt & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{dt^3}{3} & \frac{dt^2}{2}t & 0 & 0 \\ 0 & 0 & \frac{dt^2}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{dt^3}{3} & \frac{dt^2}{2} \\ 0 & 0 & 0 & 0 & \frac{dt^2}{2} & dt \end{bmatrix} \tag{5}$$

which is defined in [1]. The variable q is defined based on an understanding of the target's capabilities. Large values are used for air targets while ground targets would use smaller values.

The update equations are given as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^T + \mathbf{R}\right)^{-1} \tag{6}$$

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{K}_k\left(\mathbf{z}_k - \mathbf{h}\left(\mathbf{x}_{k|k-1}\right)\right) \tag{7}$$

$$\mathbf{P}_{k|k} = \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\right)\mathbf{P}_{k|k} \tag{8}$$

The Jacobian **H** is a linearized version of the sensor dynamics **h**(•) in relation to the target states. The three main sensor types are those that provide azimuth/elevation (angle-only measurements)

$$\mathbf{h}\left(\mathbf{x}_{k|k-1}\right) = \begin{bmatrix} \alpha \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \arctan\left(\dfrac{\left(x_{tgt} - x_{platform}\right)}{\left(y_{tgt} - y_{platform}\right)}\right) \\ \arctan\left(\dfrac{\left(z_{tgt} - z_{platform}\right)}{\sqrt{\left(x_{tgt} - x_{platform}\right)^2 + \left(y_{tgt} - y_{platform}\right)^2}}\right) \end{bmatrix} \tag{9}$$

range/bearing/elevation

$$\mathbf{h}\left(\mathbf{x}_{k|k-1}\right) = \begin{bmatrix} \rho \\ \alpha \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sqrt{\left(x_{tgt} - x_{platform}\right)^2 + \left(y_{tgt} - y_{platform}\right)^2 + \left(z_{tgt} - z_{platform}\right)^2} \\ \alpha \\ \varepsilon \end{bmatrix} \tag{10}$$

and range/bearing/elevation/range-rate

$$\mathbf{h}\left(\mathbf{x}_{k|k-1}\right) = \begin{bmatrix} \rho \\ \alpha \\ \varepsilon \\ \dot{\rho} \end{bmatrix} = \begin{bmatrix} \rho \\ \alpha \\ \varepsilon \\ \dfrac{x}{\rho}\dot{x} + \dfrac{y}{\rho}\dot{y} + \dfrac{z}{\rho}\dot{z} \end{bmatrix} \tag{11}$$

The measurement error covariance **R** is based on the sensor accuracy for the measurement **z**.

For the control law approach, the primary difference is in the prediction equations. The standard equations of the EKF with a controller become

$$\mathbf{x}_{k+1|k} = \mathbf{F}\mathbf{x}_{k|k} + \mathbf{g}\left(\mathbf{u}_k\right) \tag{12}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}\mathbf{P}_{k|k}\mathbf{F}^T + \mathbf{G}E\left[\mathbf{u}\mathbf{u}^T\right]\mathbf{G}^T + \mathbf{Q}_k \tag{13}$$

This incorporates an external input which as seen in Figure 2 allows us to incorporate a control law into the estimation effort. The new state vector becomes

$$\mathbf{x}^T = \begin{bmatrix} x & y & z \end{bmatrix} \tag{14}$$

with a kinematic equation

$$\mathbf{F} = \mathbf{I}_{3x3} \tag{15}$$

The control law becomes the velocity component of the estimation problem. If the range-rate is measurement it can be used to drive the position estimate.

3 Target Tracking Examples

Two basic target tracking problems are used to demonstrate the capabilities of the basic approach of this estimation routine.

3.1 Straight-Line Target Using Range-Bearing Measurements

The first target problem is a straight-line target with constant velocity. Figure 3 shows the target motion (dashed line) along with the sensor platform and its trajectory (solid dot line). A subset of the measurements is seen as solid lines with a diamonds on the end. The measurements are range-bearing measurements. The range error is assumed to be 1m, and the bearing error is assumed to 0.1 radians. The process noise factor, *q*, is set to 0.17 [1]. The time between measurements is 1.0 seconds.

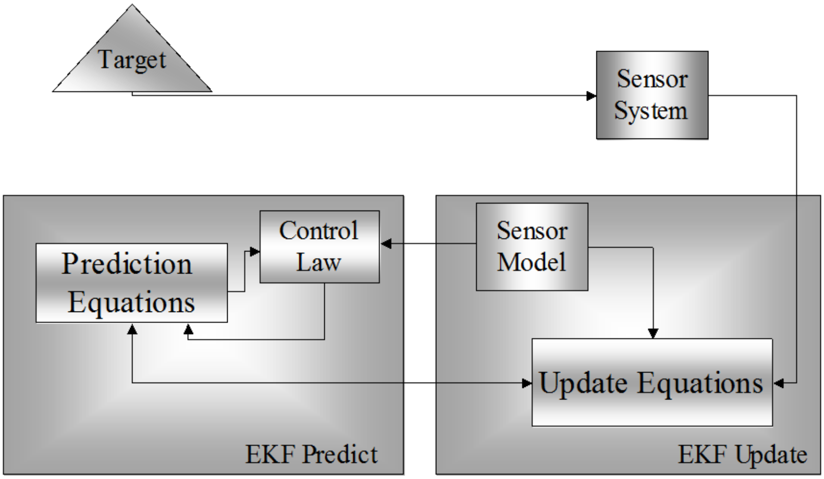


Fig. 2. With the control law, the EKF incorporates information in manners that can be much different than those of the standard estimator

The control law will use the variation in the previous updated estimate and the conversion of the measurement to the position state coordinate frame (the inverse function of state-to-measurement) and divide by the time difference between the last measurement and the current measurement:

$$\begin{bmatrix} \rho_{k+1}\sin\eta_{k+1} \\ \rho_{k+1}\cos\eta_{k+1} \end{bmatrix}\frac{1}{dt} - \mathbf{x}_{k|k} \tag{16}$$

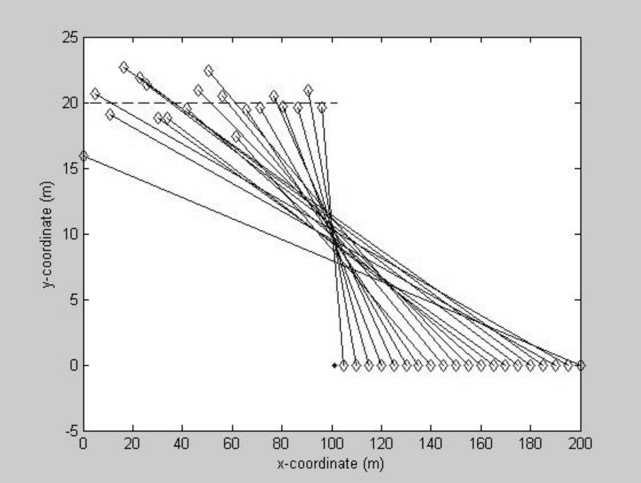


Fig. 3. The straight-line motion target. The sensor platform is lower right headed to the left with the target tracked north and west.

3.2 Maneuvering Target Using Range-Bearing-Range-Rate Measurements

The second example is that of target performing a straight-line target that is followed by a turn which the target again starts a straight-line leg as seen in Figure 4. Again, the target is denoted in the figure as the dash line. The platform is denoted by the solid-dot line. Again, the measurements are seen as solid lines with diamonds. The measurement will have a range-rate component as well with a measurement error of 0.1m/s. Otherwise, the sensor is the same. Based on implementation knowledge [4], the Jacobian for the range rate will have its position components zeroed out. The other errors and update are the same as mention in Section 3.1.

The control law uses the range rate to calculate a weight between the velocity created by the estimates and the velocity created by the estimated position and the position created by the next measurement

$$\begin{aligned} \mathbf{u}_1 &= \left(\begin{bmatrix} \rho_{k+1}\sin\eta_{k+1} \\ \rho_{k+1}\cos\eta_{k+1} \end{bmatrix} - \mathbf{x}_{k|k} \right)\frac{1}{dt} \\ \mathbf{u}_{21} &= \frac{\mathbf{x}_{k+1|k} - \mathbf{x}_{k|k-1}}{dt} \\ \mathbf{u} &= {(w_1\mathbf{u}_1 + w_2\mathbf{u}_2)}\big/{(w_1 + w_2)} \end{aligned} \tag{17}$$

where

$$
\begin{aligned}
w_1 = \Bigg| & \frac{\rho_{k+1}\sin\eta_{k+1}}{\left(\left(\rho_{k+1}\sin\eta_{k+1} - x_{1,ownship}\right) + \left(\rho_{k+1}\cos\eta_{k+1} - x_{2,ownship}\right)\right)^{1/2}} u_{1,1} + \\
& \left(\frac{\rho_{k+1}\cos\eta_{k+1}}{\left(\left(\rho_{k+1}\sin\eta_{k+1} - x_{1,ownship}\right) + \left(\rho_{k+1}\cos\eta_{k+1} - x_{2,ownship}\right)\right)^{1/2}} u_{1,2}\right) - \dot{\rho}_{k+1}\Bigg|
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
w_2 = \Bigg| & \frac{x_{1,k+1|k}}{\left(\left(x_{1,k+1|k} - x_{1,ownship}\right) + \left(x_{2,k+1|k} - x_{2,ownship}\right)\right)^{1/2}} u_{2,1} + \\
& \left(\frac{x_{2,k+1|k}}{\left(\left(x_{1,k+1|k} - x_{1,ownship}\right) + \left(x_{2,k+1|k} - x_{2,ownship}\right)\right)^{1/2}} u_{2,2}\right) - \dot{\rho}_{k+1}\Bigg|
\end{aligned}
\tag{19}
$$

4 Results

Both target tracking examples were processed through the standard EKF and the control version. A single run for each using the same noisy data was applied. The goal of this effort was to determine if a control-law-based approach could be made to provide similar results to that the standard EKF approach that uses a velocity component in the state vector. The results were compared against truth. In Figure 5, the absolute position-error was generated for both techniques applied to the nonmaneuvering target. The control approach error is denoted as the dashed line while the standard approach is shown as the solid line. The results are slightly offset but both appear to be well within the combined range and cross-range errors.

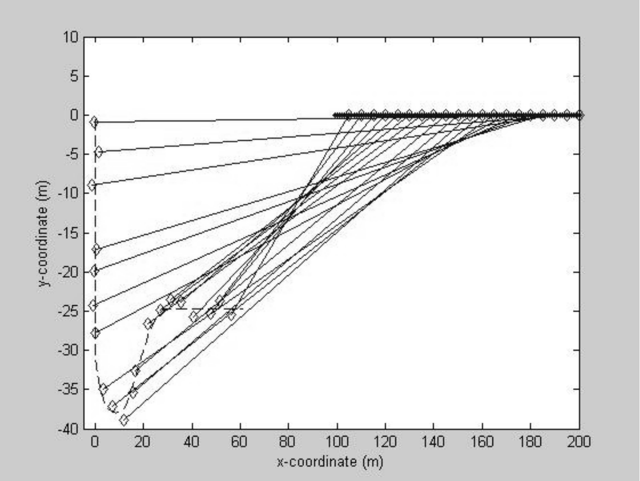


Fig. 4. The maneuver target example. The sensor platform is in the upper right with the target pulling a maneuver.

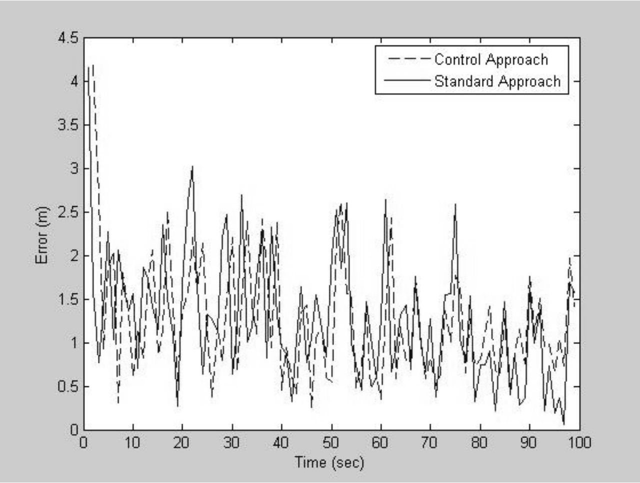


Fig. 5. Absolute position-error using both estimation techniques show similar results for the straight-line target tracking problem

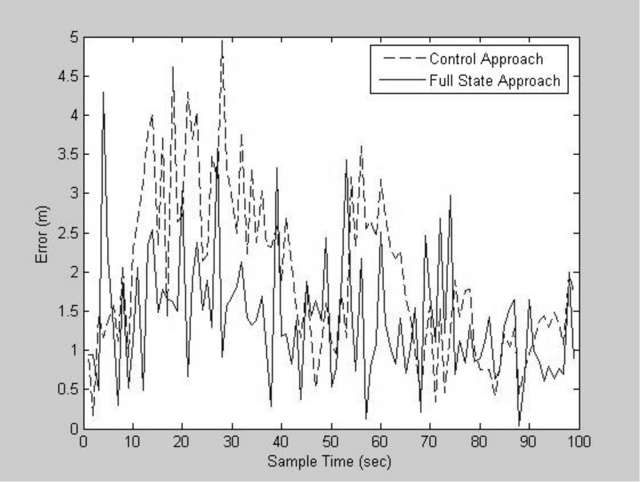


Fig. 6. Absolute position-error using both estimation techniques for the maneuvering target case shows that the control technique performs slightly worse until near the end of the scenario

In Figure 6, the absolute position-error of the tracking problem with the maneuvering target is shown. Here, the control-law approach performs slightly worst in this single case than the standard EKF approach. The control law approach still remains within the combined measurement errors (range and cross-range). Although this analysis has shown that the standard approach is better, it a single case and also indicate that the control approach will provide a comparable result in the cases where the tracker has fully observable measurements.

5 Conclusions and Future Directions

The first step in this research has shown that a control-law estimation technique can be used in a fully observable tracking problem. This allows the estimation routine to be general enough that it can be a generic tracking system. When the new algorithms for angle-only tracking and range-only tracking, the approach does not have to be used in an ad hoc manner where the technique is transitioned to when the issues with tracking arise. Analysis of this approach will be continued with Monte Carlo runs of these and different target trajectories.

In the next step of this research effort, a new control law will be developed for the angle-only tracking problem. This will be followed by a fuzzy-control algorithm that can switch between different tracking issues seamlessly.

References

1. Bar-Shalom, Y., Li, X.-R.: Estimation and Tracking: Principles, Techniques, and Software. Artech House Inc., Norwood (1993)
2. Blackman, S.: Multiple-Target Tracking with Radar Applications. Artech House, Norwood (1986)
3. Blackman, S., Popoli, R.: Design and Analysis of Modern Tracking Systems. Artech House, Norwood (1999)
4. Blackman, S.: Personal Correspondence (2013)
5. Brown, R.G., Hwang, P.Y.C.: Introduction to Random Signals and Applied Kalman Filters, 4th edn. Wiley, New York (2012)
6. Cappe, O., Godsill, S., Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of IEEE 95(5), 899 (2007)
7. Clements, R., Tavares, P., Lima, P.: Small Satellite Attitude Control Based a the Kalman Filter. In: Proceeding of the 2000 IEEE Symposium on Intelligent Control, Yokahama, Japan, pp. 79–84 (September 2000)
8. Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing 10(3), 197–208 (2000)
9. Haykin, S.: Kalman Filters and Neural Networks. John Wiley & Sons (2004)
10. Kirubarajan, T., Bar-Shalom, Y.: Tracking Evasive Move-Stop-Move Targets With A GMTI Radar Using A VS-IMM Estimator. IEEE Transactions on Aerospace and Electronic Systems 39(3), 1098–1103 (2003)
11. Kramer, K.A., Stubberud, S.C.: Tracking of multiple target types with a single neural extended Kalman filter. International Journal of Intelligent Systems 25(5), 440–459 (2010)
12. Owen, M.W., Stubberud, A.R.: NEKF IMM Tracking Algorithm. In: Drummond, O. (ed.) Proceedings of SPIE: Signal and Data Processing of Small Targets 2003, San Diego, California, vol. 5204. TBD (August 2003)
13. Stubberud, A.R.: Constrained Estimate of the State of a Time-Variable System. In: Proceedings of the 14th International Conference on Systems Engineering, Coventry, UK, pp. 513–518 (September 2000)
14. Yang, C., Blasch, E.: Fusion of Tracks with Road Constraints. Journal of Advances of Information Fusion 3(1), 14–31 (2008)

LQ Optimal Control of Periodic Review Perishable Inventories with Transportation Losses

Piotr Lesniewski and Andrzej Bartoszewicz

Institute of Automatic Control, Technical University of Lodz, 90-924 Lodz,
18/22 Bohdana Stefanowskiego St., Poland
piotr.lesniewski2@gmail.com, andrzej.bartoszewicz@p.lodz.pl

Abstract. In this paper an LQ optimal warehouse management strategy is proposed. The strategy not only explicitly takes into account decay of commodities stored in the warehouse (perishable inventory) but it also accounts for transportation losses which take place on the way from supplier to the warehouse. The proposed strategy ensures full customers' demand satisfaction and prevents from exceeding the warehouse capacity. Moreover, it guarantees that the ordered quantities of goods are bounded and it helps achieve good trade-off between fast reaction of the system to time-varying demand and the big volume of the ordered goods. These favourable properties of the proposed strategy are formally stated as three theorems and proved in the paper.

Keywords: LQ optimal control, discrete time systems, inventory control, perishable inventory.

1 Introduction

The control theoretic approach to the issue of supply chain management has recently become an important research subject. An overview of the techniques used in the field and the obtained results can be found in [1-4]. The first application of the control theory methods to the management of logistic processes was reported in the early 1950s when Simon [5] applied servomechanism control algorithm to find an efficient strategy of goods replenishment in continuous time, single product inventory control systems. A few years later the discrete time servomechanism control algorithm for the purpose of efficient goods replenishment has been proposed [6]. Since that time numerous solutions have been presented, and therefore, further in this section we are able to mention only a few, arbitrarily selected examples of solutions proposed over the last decades. In [7] and [8] autoregressive moving average (ARMA) system structure has been applied in order to model uncertain demand. Then in [9] and [10] model predictive control of supply chain has been proposed and in [11] a robust controller for the continuous-time system with uncertain processing time and delay has been designed by minimising H_∞-norm. However, practical implementation of the strategy described in [11] requires application of numerical methods in order to obtain the control law parameters, which limits its analytical tractability.

In [12] lead-time delay for conventional, non-deteriorating inventories is explicitly taken into account and represented by additional state variables in the state space description. This approach results in the optimal controller designed by minimisation of

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_5, © Springer International Publishing Switzerland 2014

quadratic performance index. An extension of the results presented in [12] to the case of perishable inventories is given in [13]. However, the analysis given in [13] does not take into account transportation losses (or in other words goods decay during the order procurement time). Therefore, in this paper we consider perishable inventories and we explicitly account for the ordered goods losses during the non-negligible lead time.

In this paper we consider a periodic-review inventory system with perishable goods replenished from a single supply source. Contrary to the previously published results we consider not only losses which take place when the commodity is stored in the warehouse, but also those which happen during the supply process, i.e. the losses on the way from the supplier to the warehouse. We propose a discrete time representation of the supply chain dynamics and we apply minimization of the quadratic performance index to design the controller for the considered system. This index takes into account not only the stock level, but also the amount of goods en route to the distribution center, which has not been considered in earlier works. The controller is determined analytically in a closed form, which allows us to state and formally prove important properties of proposed inventory policy. First, we prove that the designed management policy always generates strictly positive and upper bounded order quantities, which is an important issue from the practical point of view. Next, we define the warehouse capacity which provides enough space for all incoming shipments, and finally we derive a condition which must be satisfied in order to guarantee 100% service level, i.e. full satisfaction of imposed demand.

2 Inventory Replenishment System Model

In this paper we consider a periodic review inventory replenishment system with an unknown, time-varying demand $d(kT)$ and transportation losses. The inventory is replenished from a distant supply source with a lead time L. The lead time is a multiple of the review period T, i.e. $L = mT$, where m is a positive integer. The model of the inventory replenishment system considered in this paper is illustrated in Figure 1. The amount of goods ordered at time kT (where $k = 0, 1, 2, ...$) is denoted by $u(kT)$. The orders are determined using the current stock level $y(kT)$, the demand stock level y_d and the order history. Since we explicitly take into account transportation losses, only αu of goods (where $0 < \alpha \leq 1$) reach the warehouse. Furthermore, as we consider perishable commodities, during each review period a fraction σ $(0 \leq \sigma < 1)$ of the stock deteriorates. The demand is modeled as an *a priori* unknown, nonnegative function of time $d(kT)$, its upper bound is denoted by d_{max}. If there are enough items in stock, the demand is fully covered, otherwise, only a part of the demand is satisfied. Therefore, we introduce an additional function $h(kT)$ which represents the amount of goods actually sold at review period k. Thus

$$0 \leq h(kT) \leq d(kT) \leq d_{max}. \tag{1}$$

The current stock level can be presented in the following form

$$y[(k+1)T] = py(kT) + \alpha u[(k-m)T] - h(kT). \tag{2}$$

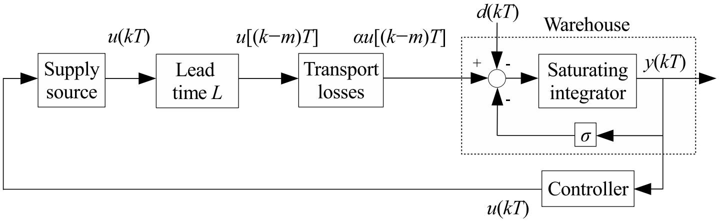


Fig. 1. Inventory system model

where $p = 1 - \sigma$ is the fraction of the stock remaining in the warehouse. Of course, since $0 \le \sigma < 1$ we have $0 < p \le 1$. We assume that the warehouse is initially empty, $y(0) = 0$, and the first order is placed at $k = 0$. The first order arrives at the warehouse at mT, and $y(kT) = 0$ for $k \le m$. We assume, that the goods (apart from the fraction $1 - \alpha$ which is broken during transport) reach the stock new, and deteriorate while kept in it. The stock level can be rewritten in the following form

$$y(kT) = \alpha \sum_{j=0}^{k-m-1} p^{k-m-1-j} u(jT) - \sum_{j=0}^{k-1} p^{k-1-j} h(jT). \tag{3}$$

In order to apply a control theoretic approach to this problem it is useful to represent the model in the state space. We select the first variable as the stock level, $x_1(kT) = y(kT)$. The remaining state variables represent the delayed values of the control signal, i.e. $x_j(kT) = u(k - n + j - 1)$ for $j = 2, \ldots, n$, where n is the system order. We can now describe the system in the state space as

$$\begin{aligned} \boldsymbol{x}[(k+1)T] &= \boldsymbol{A}\boldsymbol{x}(kT) + \boldsymbol{b}u(kT) + \boldsymbol{o}h(kT) \\ y(kT) &= \boldsymbol{q}^T\boldsymbol{x}(kT), \end{aligned} \tag{4}$$

where $\boldsymbol{A}$ is a n x n state matrix and $\boldsymbol{b}$, $\boldsymbol{o}$, $\boldsymbol{q}$ are n x 1 vectors

$$\boldsymbol{A} = \begin{bmatrix} p\ \alpha\ 0 & & 0 \\ 0\ 0\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0\ 0\ 0 & \cdots & 1 \\ 0\ 0\ 0 & & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{o} = \begin{bmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{q} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{5}$$

Since the goods perish at rate σ when kept in the warehouse, in order to keep the stock at the demand level y_d it is necessary to constantly refill it at rate σy_d. Taking into account transport losses the desired system state is therefore given by

$$\boldsymbol{x_d} = y_d \begin{bmatrix} 1 & \sigma/\alpha & \ldots & \sigma/\alpha \end{bmatrix}^T. \tag{6}$$

3 Proposed Supply Management Strategy

In this section we will develop a LQ optimal controller for the considered inventory system. Its important properties will then be formulated and proved.

In optimization problems we often consider a quadratic quality criterion that involves the control signal and the state error vector $\boldsymbol{e} = \boldsymbol{x_d} - \boldsymbol{x}$. Also in this paper we seek for a control law that minimizes the following cost functional

$$J(u) = \sum_{k=0}^{\infty} \left[u^2(kT) + \boldsymbol{e}^T(kT)\boldsymbol{W}\boldsymbol{e}(kT)\right]. \tag{7}$$

We choose $\boldsymbol{W} = \mathrm{diag}(w_1, w_2, \ldots, w_2)$ where w_1 and w_2 are positive coefficients adjusting the influence of the stock level error and the amount of goods in transit respectively. This quality criterion is more general than the one presented in [13], which simply ignored the amount of goods currently held in transport. According to [14] the optimal control $u_{opt}(kT)$ that minimizes the cost functional (7) can be presented as

$$u_{opt}(kT) = r - \boldsymbol{g}\boldsymbol{x}(kT), \tag{8}$$

where

$$\boldsymbol{g} = \boldsymbol{b}^*\boldsymbol{K}\left(\boldsymbol{I_n} + \boldsymbol{b}\boldsymbol{b}^*\boldsymbol{K}\right)^{-1}\boldsymbol{A}, \tag{9}$$

and r is a constant term. Operator $(.)^*$ denotes the complex conjugate matrix transpose, semipositive matrix $\boldsymbol{K}$ satisfies $\boldsymbol{K}^* = \boldsymbol{K}$ and is determined by the following Ricatti equation

$$\boldsymbol{K} = \boldsymbol{A}^*\boldsymbol{K}\left(\boldsymbol{I_n} + \boldsymbol{b}\boldsymbol{b}^*\boldsymbol{K}\right)^{-1}\boldsymbol{A} + \boldsymbol{W}. \tag{10}$$

Because all elements of $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{q}$ are real numbers, the complex conjugate matrix transpose $(.)^*$ is equivalent to the matrix transpose $(.)^T$. This means, that all elements of matrix $\boldsymbol{K}$ are also real numbers. Therefore, condition $\boldsymbol{K}^* = \boldsymbol{K}$ implies that $\boldsymbol{K}$ is symmetric. Because of this, in order to make notation as concise as possible, we will represent the elements of $\boldsymbol{K}$ below the main diagonal by '$*$'.

As the system order n depends on the transport delay, in order to draw general conclusions we need to solve (10) analytically for an arbitrary system order. The approach proposed here is similar to the one used in [13] and involves iterative substitution of $\boldsymbol{K}$ into the right hand side of equation (10) and comparing with its left hand side, so that at each iteration the number of independent elements of $\boldsymbol{K}$ is reduced.

We begin by substituting the most general form of $\boldsymbol{K}$ into (10) and obtain

$$\boldsymbol{K}_1 = \begin{bmatrix} k_{11} & \frac{\alpha}{p}(k_{11} - w_1) & k_{13} & & k_{1n} \\ * & \frac{\alpha^2}{p^2}(k_{11} - w_1) + w_2 & k_{23} & \cdots & k_{2n} \\ * & * & k_{33} & & k_{3n} \\ & \vdots & & \ddots & \vdots \\ * & * & * & & k_{nn} \end{bmatrix}. \tag{11}$$

The next step involves substituting (11) into (10) and comparing the left and right hand sides. We then repeat this procedure, until all elements of $\boldsymbol{K}$ are expressed as functions of k_{11}, system parameters p and α, system order n and the weighting factors w_1 and w_2

$$\boldsymbol{K} = \begin{bmatrix} k_{11} & \dfrac{\alpha}{p}(k_{11} - w_1) & & \dfrac{\alpha}{p^{n-1}}\left(k_{11} - w_1 \sum\limits_{i=0}^{n-2} p^{2i}\right) \\ * & \dfrac{\alpha^2}{p^2}(k_{11} - w_1) + w_2 & \cdots & \dfrac{\alpha^2}{p^n}\left(k_{11} - w_1 \sum\limits_{i=0}^{n-2} p^{2i}\right) \\ & \vdots & \ddots & \vdots \\ * & * & & \dfrac{\alpha^2}{p^{2n-2}}\left(k_{11} - w_1 \sum\limits_{i=0}^{n-2} p^{2i}\right) + (n-1)w_2 \end{bmatrix}. \tag{12}$$

Now we can determine k_{11} by substituting (12) into (10) and comparing the first elements of the obtained matrices. This results in

$$\begin{aligned} &k_{11}\left\{w_1\alpha^2\left(2 - p^{2n-2} + 2p^2\frac{1-p^{2n-2}}{1-p^2}\right) - p^{2n-2}(1-p^2)[1+(n-1)w_2]\right\} + \\ &- \frac{w_1^2\alpha^2(1 - p^{2n} - p^{2n-2} + p^{4n-2})}{(1-p^2)^2} + p^{2n-2}w_1[1+(n-1)w_2] - \alpha^2 k_{11}^2 = 0. \end{aligned} \tag{13}$$

Equation (13) has two roots

$$k_{11}^{\pm} = \frac{w_1\alpha^2\left(\dfrac{2 - p^{2n} - p^{2n-2}}{1-p^2}\right) - p^{2n-2}(1-p^2)[1+(n-1)w_2] \pm p^{2n-2}\sqrt{\Delta}}{2\alpha^2}, \tag{14}$$

where

$$\Delta = w_1^2\alpha^4 + 2\alpha^2 w_1[1+(n-1)w_2](p^2+1) + (1-p^2)^2[1+(n-1)w_2]^2. \tag{15}$$

Only k_{11}^{+} guarantees that $\boldsymbol{K}$ is semipositive definite. Having found $\boldsymbol{K}$, using k_{11}^{+} with (9) we derive vector $\boldsymbol{g}$

$$\boldsymbol{g} = \gamma\left[1/\alpha \; 1/p \; 1/p^2 \; \ldots \; 1/p^{n-1}\right], \tag{16}$$

where

$$\gamma = p^n\left\{1 - \frac{2[1+(n-1)w_2]}{w_1\alpha^2 + [1+(n-1)w_2](1+p^2) + \sqrt{\Delta}}\right\}. \tag{17}$$

We can now observe, that in order for the state to reach $\boldsymbol{x_d}$ defined by (6)

$$r = y_d\left[1 - p + \gamma p^{-(n-1)}\right]/\alpha. \tag{18}$$

As all state variables except x_1 are the delayed values of the control signal, we conclude, that the control signal of the LQ optimal controller is given by

$$u_{opt}(kT) = r - \frac{\gamma y(kT)}{\alpha} - \gamma p^{-n}\sum_{i=k-m}^{k-1} p^{k-i}u(iT). \tag{19}$$

This completes the design of the LQ optimal controller. Next we will present and prove important properties of the proposed control strategy.

3.1 Stability Analysis

The design method applied in this work ensures stability of the closed loop system. In order to verify this property let us notice that a discrete time system is asymptotically stable if all the roots of its closed loop state matrix $\boldsymbol{A_c}$ lie inside the unit circle on the z-plane. In our case $\boldsymbol{A_c} = \boldsymbol{A} - \boldsymbol{bg}$, and the characteristic polynomial has the following form

$$det(z\boldsymbol{I_n} - \boldsymbol{A_c}) = z^{n-1}[z - p(1 - \gamma p^{-n})]. \tag{20}$$

All roots of (20) are located inside the unit circle if $-1 < p(1 - \gamma p^{-n}) < 1$. It can be seen from (17) that $0 < \gamma < p^n$. Since $0 < p \leq 1$, we conclude, that all but one roots of (20) lie in the origin of the z-plane, and one root lies between 0 and p, depending on the value of γ. Therefore, the closed loop system is stable and no oscillations appear at the output.

Remark 1. Weighting factors w_1 and w_2 can be tuned to meet specific requirements. When $w_1 \to 0$ the value of the control signal dominates the quality criterion, and gain γ drops to zero. When $w_1 \to \infty$ for any finite w_2 the output error is to be reduced to zero as quickly as possible, no matter the value of the control signal. The controller then becomes a dead-beat scheme, its gain γ approaches p^n. Errors of states $x_2, x_3, \ldots, x_n$ represent the difference between the current replenishment orders and the steady-state order $(1 - p)y_d/\alpha$. Therefore, increasing w_2 leads to decrease of γ and vice versa.

3.2 Properties of the Proposed Controller

In this section the properties of the control strategy proposed in this paper will be stated in three theorems. In the first one we will show that generated order quantities are always nonnegative and upper bounded. The second theorem will specify the warehouse capacity needed to always accommodate the incoming shipments. The last theorem will show how to select the demand stock value in order to ensure full consumer demand satisfaction.

Theorem 1. *The order quantities generated by the control strategy* (19) *are always bounded, and satisfy the following inequality*

$$u_m \leq u(kT) \leq \max(r, u_M), \tag{21}$$

where

$$u_m = \frac{r(1-p)}{1 + p(\gamma p^{-n} - 1)}, \quad u_M = \frac{r(1-p) + \gamma d_{max}/\alpha}{1 + p(\gamma p^{-n} - 1)}. \tag{22}$$

Proof. It follows from (19) *that* $u(0) = r$. *Therefore,* (21) *is satisfied for* $k = 0$. *Substituting* (3) *into* (19) *we obtain*

$$u(kT) = r - \frac{\gamma}{\alpha}\left[\alpha \sum_{j=0}^{k-1} p^{k-m-1-j}u(jT) - \sum_{j=0}^{k-1} p^{k-1-j}h(jT)\right]. \tag{23}$$

Now we assume, that (21) *holds for all integers up to some* $l \geq 0$. *We will show that this implies that* (21) *is also true for* $l+1$. *We can rewrite* (23) *for* $k = l+1$ *as follows*

$$u[(l+1)T] = r - \gamma p^{-m} u(lT) + \frac{\gamma}{\alpha} h(lT) - \frac{\gamma}{\alpha}\left[\alpha \sum_{j=0}^{l-1} p^{l-m-j} u(jT) - \sum_{j=0}^{l} p^{l-j} h(jT)\right]. \tag{24}$$

The last term in the above equation is equal to $p[r - u(l)]$. *Consequently*

$$u[(l+1)T] = r(1-p) + p(1-\gamma p^{-n})u(lT) + \gamma h(lT)/\alpha. \tag{25}$$

Because $h(kT)$ *is always nonnegative we can obtain from* (25) *the minimum value of the control signal*

$$u(lT) \geq r(1-p)\Big/[1 + p(\gamma p^{-n} - 1)], \tag{26}$$

which shows that the first inequality in (21) *actually holds. Since* $h(kT) \leq d_{max}$ *for any* $k \geq 0$ *we can calculate the maximum value of the control signal from* (25) *as*

$$u[(l+1)T] \leq r(1-p) + p(1-\gamma p^{-n})u(lT) + \gamma d_{max}/\alpha. \tag{27}$$

First we will consider the case when $r \geq u_M$. *From* (22) *we obtain*

$$\gamma d_{max}/\alpha \leq \gamma p^{-m} r. \tag{28}$$

Using this relation with (27) *we arrive at*

$$u[(l+1)T] \leq r(1-p) + p(1-\gamma p^{-n})r + \frac{\gamma}{\alpha} d_{max} = r - \gamma p^{-m} r + \frac{\gamma}{\alpha} d_{max} \leq r. \tag{29}$$

For the second case, when $r < u_M$ *from* (25) *we get*

$$u[(l+1)T] \leq r(1-p) + p(1-\gamma p^{-n}) \frac{r(1-p) + \gamma d_{max}/\alpha}{1 + p(\gamma p^{-n} - 1)} + \gamma d_{max}/\alpha = u_M. \tag{30}$$

Taking into account relations (26), (29) *and* (30) *and using the principle of mathematical induction we conclude, that* (21) *indeed holds for any* $k \geq 0$. *This ends the proof.*

In real inventory systems it is necessary to ensure a finite warehouse size, that will always accommodate the incoming shipments. The next theorem shows, that application of our strategy ensures that the stock level will never exceed a precisely determined, *a priori* known value.

Theorem 2. *If the proposed control strategy is applied, then the on-hand stock will never exceed its demand value, i.e. for any* $k \geq 0$

$$y(kT) \leq y_d. \tag{31}$$

Proof. The warehouse is empty for any $k \leq m = n - 1$. *Therefore, we only need to show that* (31) *holds for* $k \geq n$. *We begin by assuming, that* (31) *holds for some integer*

$l \geq n$. We will demonstrate, that this assumption implies $y[(l+1)T] \leq y_d$. Applying (23) *to* (2) *we get*

$$\begin{aligned} y[(l+1)T] =& py(lT) + \alpha r - \gamma p^{-n} \sum_{j=l-m}^{l-1} p^{l-j} h(jT) - h(lT) + \\ & - \gamma p^{-m} \left[\alpha \sum_{j=0}^{l-m-1} p^{l-m-1-j} u(jT) - \sum_{j=0}^{l-1} p^{l-1-j} h(jT) \right]. \end{aligned} \tag{32}$$

We observe from (3)*, that the terms in the square brackets are equal to the on-hand stock level in period l. Consequently*

$$y[(l+1)T] = \alpha r + py(lT)(1 - \gamma p^{-n}) - \gamma p^{-n} \sum_{j=l-m}^{l-1} p^{l-j} h(jT) - h(lT). \tag{33}$$

We assumed, that $y(lT) \leq y_d$, the amount of sold goods $h(kT)$ is always nonnegative, and r is given by (18)*. Thus, from* (33) *we can obtain*

$$y[(l+1)T] \leq y_d(1 - p + \gamma p^{-(n-1)}) + y_d(p - \gamma p^{-(n-1)}) = y_d. \tag{34}$$

We conclude, using the principle of mathematical induction, that (31) *is true for $k \geq n$. As stated before, $y(kT) = 0$ for $k < n$. Therefore* (31) *holds for any $k \geq 0$.*

It follows from Theorem 2, that if we assign a storage capacity equal to y_d at the distribution center, then all incoming shipments will be accommodated, and thus the high cost of emergency storage is eliminated. A successful inventory policy is also required to ensure high demand satisfaction. The last theorem will provide a formula for the smallest possible y_d that always ensures 100% consumer demand satisfaction.

Theorem 3. *With the application of the proposed control law, if the reference stock level satisfies inequality*

$$y_d > d_{max} \frac{1 + \gamma \sum\limits_{j=-m}^{1} p^j}{1 - p + \gamma p^{-(n-1)}}, \tag{35}$$

then for any $k \geq m+1$ the stock level is strictly positive.

Proof. *We will show that if* (35) *holds, then for any $l \geq m+1$ condition $y[(l-1)T] > 0$ implies $y(lT) > 0$. Using* (1) *with* (33) *for $l \geq m+1$ we get*

$$y[(l+1)T] \geq y_d[1 - p + \gamma p^{-(n-1)}] - d_{max} \left(1 + \gamma \sum_{j=-m}^{-1} p^j \right) > 0. \tag{36}$$

The stock level $y(kT) = 0$ for $k \leq m$. Therefore, we can obtain the stock size for $k = m+1$ from (2) *as*

$$y[(m+1)T] = \alpha r - h(mT) > 0. \tag{37}$$

Again using the principle of mathematical induction with (36) *and* (37) *we conclude, that indeed if y_d satisfies* (35)*, then the stock level is positive for any $k \geq m+1$.*

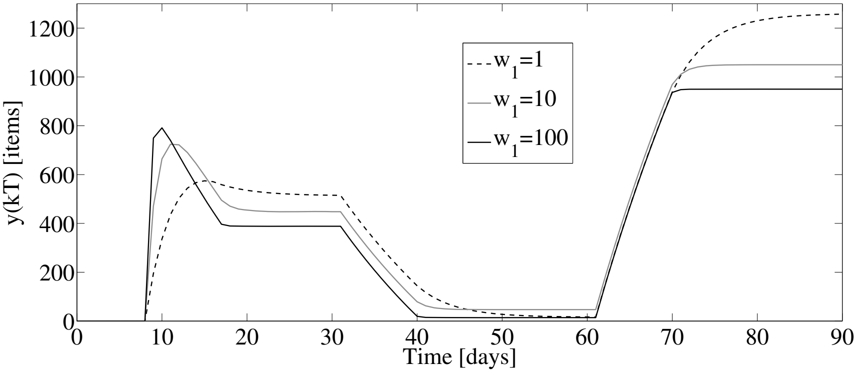


Fig. 2. Stock level for different values of w_1

We notice from (2), that a positive stock level in period k implies full satisfaction of demand in period $k-1$. Therefore, the above theorem demonstrates full satisfaction of demand $d(kT)$ for any $k \geq m$.

4 Simulation Results

In order to present the properties of the proposed control strategy, computer simulations are performed. The review period T is selected as 1 day. The lead time L is assumed to be 8 days. From this follows $m = 8$ and $n = 9$. Parameter $d_{max} = 120$ items. The actual demand is $d(kT) = 72$ for $k \in [0, 30)$, $d(kT) = 120$ for $k \in [30, 60)$ and $d(kT) = 0$ for $k \in [60, 90]$. Sudden changes of large amplitude occur in the function d, which reflects the most difficult conditions in the system. It is assumed, that 5% of the goods are broken during transport, which corresponds to $\alpha = 0.95$. The inventory decay factor $\sigma = 0.04$, which implies $p = 0.96$.

We select $w_2 = 2$ and perform three simulations, each one with a different value of w_1. The simulation parameters – gain γ obtained from (17), minimum stock demand level y'_d calculated from condition (35), and the demand stock level actually used in the simulation y_d are shown in Table 1. The results of the simulations are shown in figures 2 and 3. The value of control signal at the beginning of the transmission process is shown in Figure 4.

It can be seen from the figures, that the replenishment orders calculated by the proposed control law are always lower and upper bounded as stated in Theorem 1. Furthermore, the amount of goods in the warehouse never exceeds its demand value, and never

Table 1. Parameters of the LQ optimal controller

w_1	γ	y'_d	y_d
100	0.595	935	950
10	0.344	1031	1050
1	0.124	1244	1260

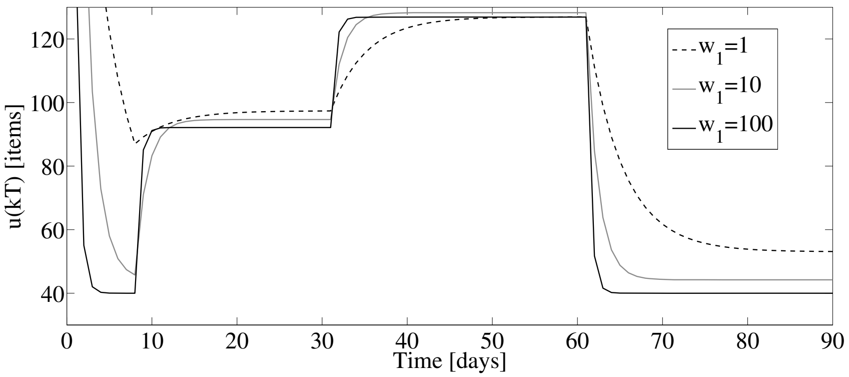


Fig. 3. Replenishment orders for different values of w_1

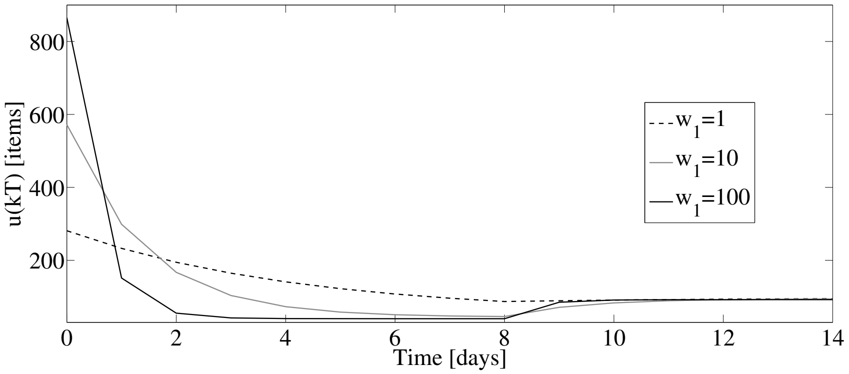


Fig. 4. Replenishment orders at the beginning of the control process

decreases to zero for $k \geq m + 1$. This means, that the incoming shipments are always accommodated in the distribution center, and that consumer demand is fully satisfied.

By selecting appropriate values of w_1 and w_2 we change the value of γ and can adapt the algorithm to particular needs. Larger values of γ result in faster tracking of consumer demand. This, in turn, allows allocating a smaller warehouse capacity, while still ensuring full consumer demand satisfaction. On the other hand, small γ leads to smaller replenishment orders at the beginning of the control process. It also makes the changes in replenishment orders smoother, which makes them easier to follow for the supplier.

5 Conclusions

In this paper an optimal periodic review supply chain management strategy has been proposed. The strategy takes into account perishable inventories with transportation losses, i.e. not only it explicitly concerns goods decay in the warehouse, but it also accounts for the losses which take place during the delivery process. The proposed strategy ensures full demand satisfaction, eliminates the risk of warehouse overflow and always generates non-negative and bounded orders. The design procedure applied in this paper is based on minimization of a quadratic cost functional (which is more general than the similar ones proposed earlier [13]), and solving the resulting matrix Ricatti equation.

Acknowledgments. This work has been performed in the framework of a project ”Optimal sliding mode control of time delay systems” financed by the National Science Centre of Poland decision number DEC 2011/01/B/ST7/02582. Kind support provided by the Foundation for Polish Science under Mistrz grant is also acknowledged.

References

1. Sarimveis, H., Patrinos, P., Tarantilis, C.D., Kiranoudis, C.T.: Dynamic modeling and control of supply chain systems: a review. Computers and Operations Research 35(11), 3530–3561 (2008)
2. Karaesmen, I., Scheller-Wolf, A., Deniz, B.: Managing perishable and aging inventories: review and future research directions. In: Kempf, K., Keskinocak, P., Uzsoy, R. (eds.) Handbook of Production Planning. Kluwer, Dordrecht (2008)
3. Boccadoro, M., Martinelli, F., Valigi, P.: Supply chain management by H-infinity control. IEEE Trans. on Automation Science and Engineering 5(4), 703–707 (2008)
4. Hoberg, K., Bradley, J.R., Thonemann, U.W.: Analyzing the effect of the inventory policy on order and inventory variability with linear control theory. European J. of Operations Research 176(3), 1620–1642 (2007)
5. Simon, H.A.: On the application of servomechanism theory in the study of production control. Econometrica 20(2), 247–268 (1952)
6. Vassian, H.J.: Application of discrete variable servo theory to inventory control. Arthur D. Little, Inc., Cambridge (1954)
7. Gaalman, G., Disney, S.M.: State space investigation of the bullwhip problem with ARMA(1,1) demand processes. International J. of Production Economics 104(2), 327–339 (2006)
8. Gaalman, G.: Bullwhip reduction for ARMA demand: the proportional order-up-to policy versus the full-state-feedback policy. Automatica 42(8), 1283–1290 (2006)
9. Aggelogiannaki, E., Doganis, P., Sarimveis, H.: An adaptive model predictive control configuration for production-inventory systems. International J. of Production Economics 114(1), 165–178 (2008)
10. Li, X., Marlin, T.E.: Robust supply chain performance via Model Predictive Control. Computers & Chemical Engineering 33(12), 2134–2143 (2009)
11. Boukas, E.K., Shi, P., Agarwal, R.K.: An application of robust technique to manufacturing systems with uncertain processing time. Optimal Control Applications and Methods 21(6), 257–268 (2000)
12. Ignaciuk, P., Bartoszewicz, A.: Linear-quadratic optimal control strategy for periodic-review inventory systems. Automatica 46(12), 1982–1993 (2010)
13. Ignaciuk, P., Bartoszewicz, A.: Linear-quadratic optimal control of periodic-review perishable inventory systems. IEEE Trans. on Control Systems Technology 20(5), 1400–1407 (2012)
14. Kwakernaak, H., Sivan, R.: Linear Optimal Control Systems. Wiley-Interscience, New York (1972)

A Dynamic Vehicular Traffic Control Using Ant Colony and Traffic Light Optimization

Mohammad Reza Jabbarpour Sattari, Hossein Malakooti, Ali Jalooli,
and Rafidah Md Noor

Faculty of Computer Science and Information Technology, University of Malaya,
50603 Kuala Lumpur, Malaysia

Abstract. Vehicle traffic congestion problem in urban areas due to increased number of vehicles has received increased attention from industries and universities researchers. This problem not also affects the human life in economic matters such as time and fuel consumption, but also affects it in health issues by increasing CO2 and greenhouse gases emissions. In this paper, a novel cellular ant-based algorithm combined with intelligent traffic lights based on streets traffic load condition has been proposed. In the proposed method road network will be divided into different cells and each vehicle will guide through the less traffic path to its destination using Ant Colony Optimization (ACO) in each cell. Moreover, a new method for traffic lights optimization is proposed in order to mitigate the traffic congestion at intersections. Two different scenarios have been performed through NS2 in order to evaluate our traffic lights optimization method. Based on obtained results, vehicles average speed, their waiting time and number of stopped vehicles at intersections are improved using our method instead of using usual traffic lights.

1 Introduction

Over the last decade, vehicle population has been increased sharply in the world. This large number of vehicles leads to a heavy traffic congestion and consequently,lots of accidents.According to RACQ Congested Roads report [1], fuel consumption, CO2 and greenhouse gases emissions, long travel time and accidents are both direct and indirect results of vehicle traffic congestion and rough (vs. smooth) driving pattern.

Accordingly, there should be a way to alleviate the vehicle congestion problem. Building new high capacity streets and highways can mitigate some of the aforementioned problems. Nevertheless, this solution is very costly,time consuming and in most of the cases, it is not possible because of the space limitations.On the other hand, optimal usage of the existent roads and streets capacity can lessen the congestion problem in large cities at the lower cost.However, this solution needs accurate information about current status of roads and streets which is a challenging task due to quick changes in vehicular networks and environments. Providing alternative paths with shortest time duration instead of shortest path distances can be useful because of lower fuel consumption and

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_6, © Springer International Publishing Switzerland 2014

traffic congestion. These approaches are called Dynamic Traffic Routing System (DTRS). Various DTRSs are proposed in [2–4], but among them, using Multi Agent System (MAS) is reported as a promising and one of the best approaches for dynamic problems [5]. Particularly, ant agents have proven to be superior to other agents in [6–8].

In ant-based algorithms, inspired from real ants behavior, artificial ants (agents) find the shortest path from source to destination based on probabilistic search in the problem space. Dividing the routing space into several smaller spaces (cells) can lead to a better routing result because of dynamic nature of the vehicle's congestion. In addition to vehicle routing and traffic control, intersections can affect the traffic congestion and smooth driving pattern. This is because of traffic lights existence in the intersections. In addition, according to [9], up to 90% of the utilized traffic lights operate based on fixed assignments of green splits and cycle duration which leads to inessential stops of vehicles. Hence, optimizing the traffic lights can eliminate waste of time and money.

Therefore, we addressed some of aforementioned drawbacks by proposing a cellular ant-based algorithm applied to dynamic traffic routing , using optimized traffic lights for traffic congestion problem in vehicular environment. The rest of this paper in organized as follows: Section 2 discusses about related works in two different sections, dynamic vehicle routing using ACO and traffic lights optimization. Proposed methods for vehicle routing and traffic lights optimization are explained in Section 3. Obtained results are discussed and justified in Section 4. Section 5 concludes the paper.

2 Related Work

In this section most of the related approaches to our topic will be discussed. Since our approach has two parts, this section is divided into two subsections; dynamic traffic routing using ACO and traffic lights optimization. To the best of our knowledge, there is no approach which utilizes both of these approaches at the same time to reduce vehicle traffic congestion and our approach uses these two methods simultaneously for first time.

2.1 Traffic Light Optimization (TLO)

During last four decades, Urban Traffic Control (UTC) based on traffic light optimization has been attracted researchers and industries attention. Complex mathematical formulas and models are used in most of the existing UTC approaches in order to traffic lights optimization. SCATS (Sydney Coordinated Adaptive Traffic System) [10] and SCOOT (Split, Cycle and Offset Optimization Technique) [11] are the most well-known examples of this kind of UTC systems. adaptive traffic light approach based on wireless sensors is proposed in [12–14]. As compared to UTC system, by using these approaches more information such as vehicles direction, speed and location can be used for getting accurate decisions for TLO. Therefore, UTC systems problem which comes from the fixed location of the detectors is solved in adaptive traffic light algorithms. In-vehicle Virtual Traffic Light (VTL) protocol is designed in [15] in order to traffic flow optimization at intersection without using road side infrastructure such as RSUs and traffic lights.

2.2 Dynamic Traffic Routing (DTR) using Ant Colony Optimization

Over dynamically changing networks, finding best routes can also be fulfilled through using Swarm Intelligence (SI) based methods. One of the most advantageous SI methods for exploring optimal solutions at low computational cost is ant routing algorithm. AntNet is a routing algorithm which is inspired by the natural ants behavior and operates based on distributed agents [8]. AntNet has been proved to be an adoptable algorithm to the changes in traffic flows and have better performance than other shortest path algorithms [7].Using ant colony algorithm in combination with network clustering autonomous system has been proved to be effective in finding best routing solutions by Kassabalidis et al. in [16]. Cooperation among neighboring nodes can be increased using a new type of helping ants which are introduced in [17]. Consequently, AntNet algorithms convergence time will be reduced as well. A new version of AntNet algorithm which improves the average delay and the throughput is introduced by Tekiner et al. in [18]. Moreover, the ant/packet ratio is used in their algorithm to constrain the number of using ants.

In road traffic routing, the significant role of Dynamic traffic routing algorithms to prevent facing congestion offer better routes to cars is noticeable. For car navigation in a city, a DTR which utilizes the Ant Based Control algorithm (ABC algorithm) is introduced in [2]. However, it is proved that this algorithm is more appropriate in small networks of city streets rather than big ones due to its scalability problems. An adjustment to the AntNet and the Ant Based Control (ABC) to direct drivers to the best routs by the aid of historically-based traffic information has been offered in [3].Another version of the AntNet algorithm by the help of which travel time can be improved over a congested network is presented in [19] and [20]. This improvement can be achieved through diverting traffic from congested routs.In hierarchical routing system (HRS), which is proposed in [4], roads are assigned to different hierarchy levels and consequently, a traffic network is split into several smaller networks or sectors. A routing table is for leading the cars to better routes is located at the networks intersections (at sector level and locally). For dynamic routing, an ant-based algorithm is utilized. The high adaptability of this approach in complex networks is noticeable.

3 Proposed Model

3.1 VANET based Traffic Light Optimization

In this section, we propose a new vehicle-to-traffic light counter based model which is used for traffic lights optimization as well as finding optimal path for

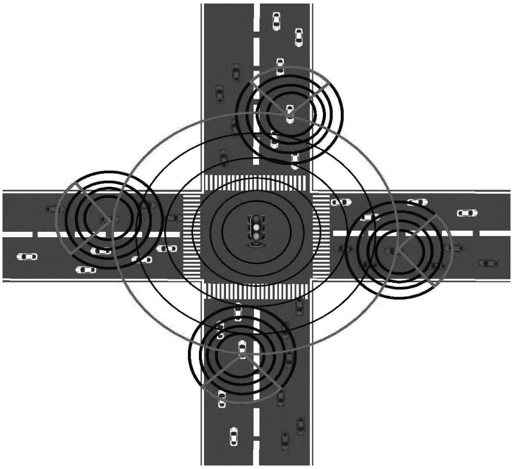


Fig. 1. Traffic light optimization model

vehicles. Traffic lights optimization means that different green and red light duration will be assigned to different streets in an intersection by intelligent traffic light instead of fixed and predefined duration. Optimal path is used for a path with low traffic and reasonable distance to destination in this paper. This model is illustrated in Figure 1 and is called Wings because of its similarity to real wings.

Referring to Figure 1, wireless devices are mounted on traffic light and vehicles. Consequently, vehicles inside the communication range of traffic lights, will send a message to their front traffic light. This message contains vehicles location, speed, direction and Received Signal Strength (RSS). Traffic lights will determine the farthest vehicle in each street based on their locations and RSS, and send request message to them. These farthest vehicles broadcast a request message in their communication rang in order to response to traffic lights request message. Vehicles located behind the farthest vehicle in this communication range will response to this message. Number of received messages will be sent to traffic light through the farthest vehicle. Thus, traffic light can calculate the number of near vehicles to intersection on ith street (N_{vi}) and total number of near vehicles to intersection (T_v) based on real-time data. A fixed and constant cycle length (C_l) is assigned to each traffic light by considering different parameters such as number of lanes, width of the streets, downtown vs. suburban streets and intersections and etc. Using this information, traffic light will assign different green light time durations (G_{ti}) to different streets in its communication range. These times will be calculated for each street by below formula:

$$G_{ti} = \frac{N_{vi}}{T_v} * C_l \quad (1)$$

Table 1 shows an example of green light time duration calculation for an intersection. Traffic light cycle length is assumed as 4 minutes.

Table 1. An example of green time duration calculation by traffic light

Road ID	Number of Vehicles	Assigned green time duration
A	14	63 sec
B	20	90 sec
C	2	9 sec
D	44	198 sec

3.2 Vehicular Routing with Optimal Path

As discussed in introduction section, through the past decade the number of vehicles grows sharply and cause many problems such as vehicles traffic and accident, long travel time, high CO2 and greenhouse gases emissions and fuel consumption. Building new high capacity streets and highways can alleviate some of aforementioned problems. However, this solution is very costly and most of the cases are not possible because of space limitations. Using Vehicle Route Guidance System (VRGS) is another way to utilize the roads capacity efficiently by proposing source-to-destination paths to drivers considering different objectives such as shortest or toll-free paths. But, most of the available navigators are using static routing algorithm such as Dijkstra or A* algorithms or in the best case are using dynamic traffic information (like TMC) but with rather high update intervals of several minutes. Nevertheless, these approaches need centralized process unit to compute the best or shortest path which limits the covered area and need high map update intervals in the system.

Thus, we propose a new decentralized routing algorithm based on real-time traffic information using vehicular networks and ant colony algorithm. Since, the vehicles traffic is the main source of problems in vehicle management systems based on RACQ Congested Roads report in [1]. Thus, our proposed algorithm is aiming to reduce the traffic in order to increase throughput while avoid to create another bottleneck at other street. Because congestion condition in vehicular networks is very dynamic and change as time goes by. Thus, we divide routing map into different cells and routing will be done based on current traffic condition on each cell. Moreover, layered model is used in order to reduce the computing overhead as well as increase the coverage area. Our layered and cellular model is illustrated in Figure 2 and is explained as follow:

Referring to Figure 2, our proposed model contains three different bottom-up layers:

1. Physical layer: This layer shows the real road map, nodes correspond to intersections, junctions, meanwhile, links correspond to streets and highways. This map can be exported from map databases like OpenStreetMap. This layer will be used for intra-cell (inside one cell) routing in our algorithm. This layers graph is given by $G_p = (N_p, L_p)$, where Np and L_p is the set of nodes and links, respectively. At each specific time (ti), a weight will be assigned to each link in the graph based on vehicles density ($NV_{ij}(_{ti})$). Vehicles density can be obtained from different tools such as Road Side Units

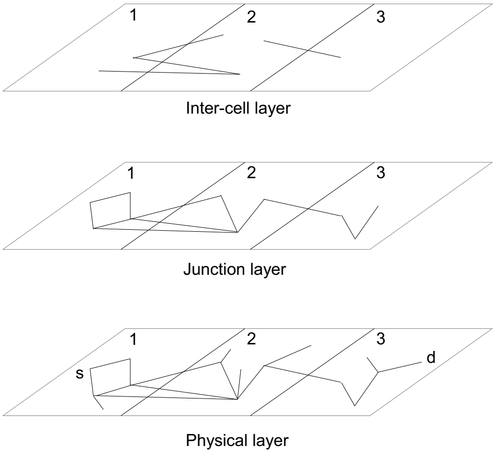


Fig. 2. Proposed layered and cellular model used in ant-based vehicle routing algorithm

(RSUs), Inductive Loop Detectors (ILD) [21] and Video Imaging Vehicle Detection System (VIVDS) [22]. In this paper, we assumed that each streets density is available through one of above mentioned ways. α_{ij} presents link weight between nodes i and j. In this paper, number of vehicles and links weight has inverse relationship, thus, α_{ij} can be calculated as follow:

$$\alpha_{ij} = \frac{1}{NV_{ij(ti)}} \tag{2}$$

2. Junction layer: In this layer, irrelevant nodes in physical layer which dont represent a junction are pruned.
3. Inter-cell layer: Junctions and their links which connect two different cells in junction layer will remain, otherwise, they are pruned. The remained nodes (junctions) are called border nodes. This layers information will be used whenever a vehicle travels over larger distances and thus traverses more than one cell to reach its destination. Inter-cell (between two different cells) routing table will be created based on this layers information. Table 2 is an example of inter-cell routing table for cell 1 of inter-cell layer illustrated in Figure 2.

Table 2. Inter-cell routing table for cell 1

Source	Destination	Vehicle Density
A1	A2	NV$_{A1,A2}(_{ti})$
A1	B2	NV$_{A1,B2}(_{ti})$
B1	B2	NV$_{B1,B2}(_{ti})$

In these tables, first and second columns show the existing path(s) between two different cells, and last column indicates vehicle density at particular time and thus it will be changes as time goes by based on number of vehicles on that path. Each cells inter-cell routing table will be disseminated among all junctions of same cell.

For example, Table 2 means that there are 3 outgoing links from 1st cell through two border nodes (A1, B1) to 2nd cell. Consequently, if a vehicle locates in cell 1 and want to travel to other cells (such as cell 2 or 3), first will be guided to one of these border junctions based on traffic condition using ACO algorithm, then based on traffic condition will be routed to one of border nodes in cell 2 using inter-cell routing table. If there are two or more path between two different cells (e.g. our example), path will lowest traffic will be selected.Therefore, vehicles will be routed through shortest low traffic paths, since researchers in [23] have been proven that ants find the shortest path.

Our last topic in this section is related to intra-cell routing process using ant-based agents. This process is based on ants behavior discussed in section 2. Our proposed algorithm contains three main steps:

1. Initialization: the pheromone values (weights), α_{ij}, on each link (path) are set based on vehicles density.
2. Pheromone Update: ants start to discover the rout between source and destination, and move to one of neighbor nodes based on pheromone values. This value will be decreased in two ways: first, over time by a factor ε using formula (3) and second, whenever an ant agent pass the link for finding a rout from source to destination by a factor β until a stop criterion (reach to destination or MAX-HOPS) is met by using formula (4).

$$\alpha_{ij}(t+1) = \alpha_{ij}(t) - \varepsilon \tag{3}$$

$$\alpha_{ij}(new) = (1-\beta) * \alpha_{ij}(current) \tag{4}$$

   This decreasing is done due to improve the exploration factor of the search. Because in this way, more new routes different from previous ones will be discovered and they can be used for traffic congestion mitigation purposes. MAX-HOPS is a constant value used for limiting the ants movements (e.g. Time To Live (TTL) value for routing packets).
3. Solution Construction: in this step the pheromone value will be increased only when an ant reaches the destination before it reaches MAX-HOPS. Ant backtracks to increase the pheromone levels on the links in found path by factor δ using formula (5).

$$\alpha_{ij}(new) = \alpha_{ij}(current) - \delta \tag{5}$$

In most other approaches, decreasing and increasing of pheromone values happen globally which requires synchronization and more communication. However, these updates are happened locally in our proposed method.

4 Simulation Results and Discussion

In order to evaluate our proposed approach for traffic lights optimization, DIVERT simulator [24] is used. A road topology with two intersections and bidirectional streets is used for evaluation. 100 vehicles with various speeds ranging from 40 km/h to 90 km/h, are distributed randomly in this topology. These vehicles moves toward an specific predefined points based on their directions Two scenarios, one with usual traffic lights and another with our proposed adaptive and dynamic traffic lights, were considered in this simulation. Vehicles average speed and waiting time as well as number of stopped vehicles are three evaluation metrics in our simulation. The number of stopped vehicles behind the traffic lights at intersections is demonstrated in Figure 3. Based in this figure during the simulation time, the number of stopped vehicles in adaptive and dynamic traffic lights (our proposed method) is less than the number of stopped vehicles

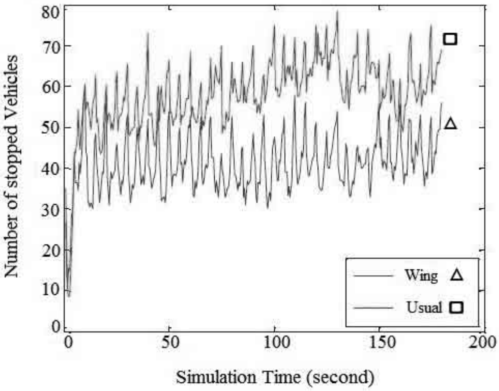


Fig. 3. Number of stopped vehicles for two scenarios

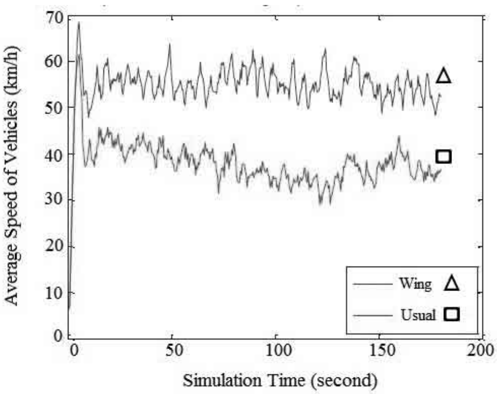


Fig. 4. Average speed of vehicles in two scenarios

in usual traffic light scenario. It means that traffic congestion at intersections are reduced using our method. Figure 4 compares the vehicles average speed in two aforementioned scenarios. Referring to Figure 4, vehicles average speed is higher in the case of using our proposed method because of lower number of stops and longer green times for high traffic streets.

5 Conclusion and Future Work

In this paper, we addressed one the most important problems in transportation system which is vehicle traffic congestion problem. Based on our literature review, traffic lights optimization and dynamic vehicle routing are two main approaches used for solving traffic congestion problem. Therefore, a cellular ant-based algorithm using optimized traffic lights which combines these two methods is proposed for traffic congestion problem in vehicular environments. Ant-based algorithm is used due to their superior ability in solving dynamic problems. Moreover, our traffic light optimization is done without using road side units which reduces the whole cost of the approach. This optimization examined through a simulation and results indicate that vehicles average speed and number of stopped vehicles at intersections are improved significantly as compared with usual traffic lights. As future work, we plan to implement the second part of our approach which is cellular ant-based algorithm for vehicle routing through shortest path with less traffic and compare it with static (e.g. Dijkstra) and dynamic (e.g. Dynamic System for the Avoidance of Traffic Jams (DSATJ)) approaches.

Acknowledgment. This research is supported by the Ministry of Higher Education (MOHE) and University of Malaya (project no. HIR-MOHEB00009).

References

1. Spalding, S.: Racq congested roads report (2008)
2. Kroon, R., Rothkrantz, L.: Dynamic vehicle routing using an abc-algorithm. Collected Papers on the PITA Project, 21 (2003)
3. Suson, A.C.: Dynamic routing using ant-based control
4. Tatomir, B., Rothkrantz, L.: Hierarchical routing in traffic using swarm-intelligence. In: Intelligent Transportation Systems Conference, ITSC 2006, pp. 230–235. IEEE (2006)
5. Kponyo, J.J., Kuang, Y., Li, Z.: Real time status collection and dynamic vehicular traffic control using ant colony optimization. In: 2012 International Conference on Computational Problem-Solving (ICCP), pp. 69–72. IEEE (2012)
6. Bonabeau, E., Dorigo, M., Theraulaz, G.: Swarm intelligence: from natural to artificial systems. Number 1. OUP USA (1999)
7. Dhillon, S., Van Mieghem, P.: Performance analysis of the antnet algorithm. Computer Networks 51(8), 2104–2125 (2007)
8. Di Caro, G., Dorigo, M.: Antnet: Distributed stigmergetic control for communications networks. arXiv preprint arXiv:1105.5449 (2011)

9. Stevanovic, A.: Adaptive traffic control systems: domestic and foreign state of practice. Number Project 20-5, Topic 40-03 (2010)
10. Akcelik, R., Besley, M., Chung, E.: An evaluation of scats master isolated control. In: Proceedings of the 19th ARRB Transport Research Conference (Transport 1998), pp. 1–24 (1998)
11. Robertson, D.I., Bretherton, R.D.: Optimizing networks of traffic signals in real time-the scoot method. IEEE Transactions on Vehicular Technology 40(1), 11–15 (1991)
12. Gradinescu, V., Gorgorin, C., Diaconescu, R., Cristea, V., Iftode, L.: Adaptive traffic lights using car-to-car communication. In: IEEE 65th Vehicular Technology Conference, VTC 2007-Spring, pp. 21–25. IEEE (2007)
13. Zhou, B., Cao, J., Zeng, X., Wu, H.: Adaptive traffic light control in wireless sensor network-based intelligent transportation system. In: 2010 IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall), pp. 1–5. IEEE (2010)
14. Faye, S., Chaudet, C., Demeure, I.: A distributed algorithm for multiple intersections adaptive traffic lights control using a wireless sensor networks. In: Proceedings of the First Workshop on Urban Networking, pp. 13–18. ACM (2012)
15. Ferreira, M., Fernandes, R., Conceição, H., Viriyasitavat, W., Tonguz, O.K.: Self-organized traffic control. In: Proceedings of the seventh ACM International Workshop on VehiculAr InterNETworking, pp. 85–90. ACM (2010)
16. Kassabalidis, I., El-Sharkawi, M., Marks, R., Arabshahi, P., Gray, A., et al.: Adaptive-sdr: Adaptive swarm-based distributed routing. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN 2002, vol. 1, pp. 351–354. IEEE (2002)
17. Soltani, A., Akbarzadeh-T, M.R., Naghibzadeh, M.: Helping ants for adaptive network routing. Journal of the Franklin Institute 343(4), 389–403 (2006)
18. Tekiner, F., Ghassemlooy, Z., Al-khayatt, S.: Antnet routing algorithm-improved version. In: CSNDSP 2004, Newcastle, UK, pp. 22–22 (2004)
19. Tatomir, B., Rothkrantz, L.: Dynamic traffic routing using ant based control. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3970–3975. IEEE (2004)
20. Claes, R., Holvoet, T.: Ant colony optimization applied to route planning using link travel time predictions. In: 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), pp. 358–365. IEEE (2011)
21. Roland, N.: Inductive loop detector. US Patent 20,050,035,880 (February 17, 2005)
22. Michalopoulos, P.G.: Vehicle detection video through image processing: the autoscope system. IEEE Transactions on Vehicular Technology 40(1), 21–29 (1991)
23. Jayadeva, Shah, S., Bhaya, A., Kothari, R., Chandra, S.: Ants find the shortest path: a mathematical proof. Swarm Intelligence 7(1), 43–62 (2013), doi: 10.1007/s11721-013-0076-9
24. LIACC: Divert: Development of inter-vehicular reliable telematics (2013)

Robust Inventory Management under Uncertain Demand and Unreliable Delivery Channels

Przemysław Ignaciuk

Institute of Information Technology, Lodz University of Technology,
215 Wólczańska St., 90-924 Łódź
przemyslaw.ignaciuk@p.lodz.pl

Abstract. The paper considers the problem of establishing an efficient supply policy for production-inventory systems in which stock replenishment process is unreliable. In the analyzed setting, the stock at a goods distribution center, used to satisfy uncertain, variable demand is refilled using multiple delivery channels. Due to information distortion, product defects, or improper transportation, the shipments may arrive damaged, or incomplete. The setting is modeled as a time-varying discrete-time system with multiple input-output delays. A new delay compensation mechanism, which provides smooth ordering pattern and ensures closed-loop stability for arbitrary delay, is proposed. Conditions for achieving full satisfaction of the *a priori* unknown demand are specified and formally proved.

Keywords: inventory control, time-delay systems, discrete-time systems.

1 Introduction

It has been argued in a number of recent works [1, 2] that in the currently observed increased competition and demand diversity one might seek performance improvements in production-inventory systems through the application of systematic design techniques. Therefore, as opposed to the classical stochastic, or heuristic solutions to inventory control problem, in this work formal approach is adopted.

In the considered class of systems the stock accumulated at a goods distribution center is used to satisfy uncertain market demand. Neither the value, nor statistics of demand are known *a priori*, and thus it is treated as a disturbance. The stock is replenished with delay using (possibly) multiple supply sources [3], or delivery channels [4]. Unlike the previous studies in a similar setting with nonnegligible delay and uncertain demand [5]–[11], in this paper, the situation when the stock replenishment process itself is subject to perturbation is analyzed. The investigated extra source of uncertainty is related to information distortion (e.g. erroneous order handling) and faults in goods production and delivery. In order to perform sound, formal analysis, a new model of the considered class of production-inventory systems is constructed. In the proposed framework, the uncertainty related to unreliable delivery channels is modeled as an external multiplicative disturbance. A new control strategy (ordering

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_7, © Springer International Publishing Switzerland 2014

policy), explicitly taking into account the problems related to unreliable replenishment process, is proposed. A pivotal role in establishing robust yet efficient ordering pattern plays a new delay compensation mechanism incorporated in the proposed control scheme. The designed compensation mechanism allows one to maintain closed-loop stability without compromising response speed, which is difficult to achieve in time-delay systems subject to perturbations [12, 13].

It is shown that the order quantities generated by the proposed strategy are always nonnegative and bounded, which is required for the practical implementation of any efficient ordering policy. It is also demonstrated that in the analyzed control system the available stock is never entirely depleted despite unpredictable demand variations and delivery channel uncertainty. As a result, all of the imposed demand can be satisfied from the readily available resources and maximum service level is obtained. Moreover, the storage space to accommodate all the incoming shipments is indicated, which helps in establishing suitable warehousing solutions. The crucial system properties are strictly proved and illustrated with numerical data.

2 System Model

The model of the analyzed production-inventory system is illustrated in Fig. 1. The system variables are inspected at regular, discrete time instants kT, where T is the review period and $k = 0, 1, 2, ...$ In order to save on notation in the remainder of the paper k will be used as the independent variable in place of kT.

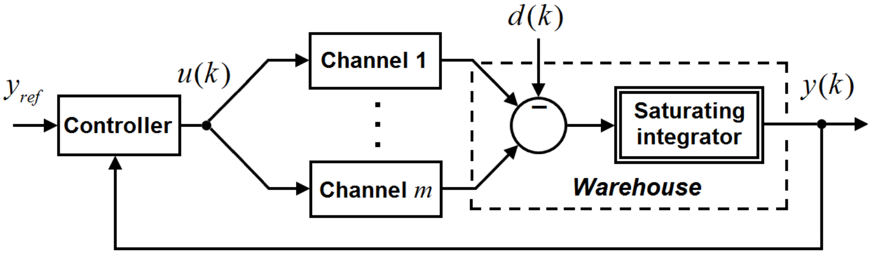


Fig. 1. Model of production-inventory system with unreliable delivery channels

The imposed demand (the goods quantity requested from inventory in period k) is modeled as an *a priori* unknown, bounded function of time $d(k)$, $0 \leq d(k) \leq d_{\max}$, where $d_{\max}$ is a positive constant denoting the estimate of maximum demand. From the control system perspective, the demand, being an exogenous, uncertain signal, is treated as a disturbance. If there is sufficient amount of goods in the warehouse to satisfy the imposed demand $d(k)$, then the actually met demand $h(k)$ (the goods sold to the customers or sent to the retailers in a distribution network) will be equal to the requested one. Otherwise, the imposed demand is satisfied only from the arriving shipments and additional demand is lost (it is assumed that the sales are not backordered and the excessive demand is equivalent to a missed business opportunity). Thus, one may write

$$0 \le h(k) \le d(k) \le d_{\max}. \tag{1}$$

The total order quantity u is calculated using the information about the current stock level $y(k)$, the stock reference value y_{ref}, and orders history. The obtained quantity is split among m delivery options (being different supply sources, or various transport and production channels) according to the company sourcing strategy. Consequently, in each review period λ_p of the total order is to be acquired using option p ($p = 1, 2, ..., m$), where λ_p is a real number from the interval [0, 1] satisfying the condition $\sum_{p=1}^{m} \lambda_p = 1$. The values of λ_p for a particular business setting are determined using a separate algorithm, e.g. [4], that can be independently incorporated in the considered framework.

The goods ordered using option p are delivered with lead-time delay L_p assumed to be a multiple of the review period, i.e. $L_p = n_pT$, where n_p is a positive integer. Hence, it is expected to receive $\lambda_p u(k - n_p)$ units of merchandise using option p in period k. However, due to mistakes in order processing, damage incurred during transportation, or product defects, only a certain part of the received shipments, $\varepsilon_p(k)$, $0 \le \varepsilon_p(k) \le 1$, is admitted to the selling process at the distribution center. The value of $\varepsilon_p(k)$ is not known *a priori*. The boundary value $\varepsilon_p = 1$ reflects the case of perfect replenishment, whereas $\varepsilon_p = 0$ means total delivery failure (e.g. due to a traffic accident, or a broken production line). Consequently, the realized order, i.e. the goods actually admitted into the distribution system acquired using option p,

$$\tilde{u}_p(k) = \varepsilon_p(k)\lambda_p u(k - n_p), \tag{2}$$

constitute an uncertain part of the order placed in period $k - n_p$. Note that similarly as in the case of demand, the definition of failure rate is general enough to accept any statistical distribution and arbitrary dynamics, i.e. no simplifying assumption concerning the rate of ε_p parameter variation is taken in the proposed model. Without loss of generality, the replenishment options may be ordered according to the associated lead time in the following way: $L_1 \le L_2 \le ... \le L_m$.

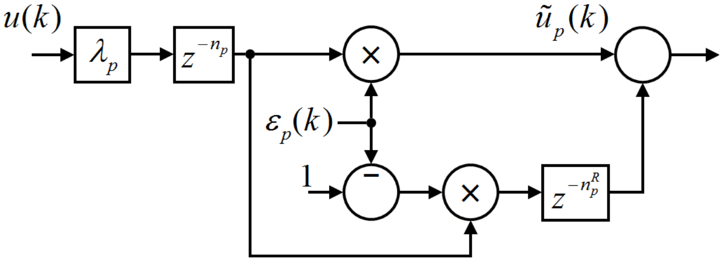


Fig. 2. Uncertain production-delivery channel

The delivery shortage $(1 - \varepsilon_p)\lambda_p u$ is supposed to be refilled (repaired) within n_p^R periods. Typically, the repair time related to option p, $n_p^R = n_p$, which means to replenish the shortage together with the currently placed order. When $n_p^R < n_p$ refilling is performed as an emergency shipment. Thus, the stock balance equation in the analyzed system takes the following form

$$\begin{aligned} y(k+1) = y(k) - h(k) &+ \sum_{p=1}^{m} \lambda_p \varepsilon_p(k) u(k - n_p) \\ &+ \sum_{p=1}^{m} \lambda_p [1 - \varepsilon_p(k - n_p^R)] u(k - n_p - n_p^R). \end{aligned} \tag{3}$$

As opposed to the systems with perfect input channels [5]–[11], in the considered case of faulty replenishment process, one needs to account for the additional delay line associated with each input-output channel. Moreover, unlike the case of perfect replenishments, the unreliable delivery channels considered in this work are characterized by a set of uncertain, time-varying parameters ε_p. Consequently, more care needs to be taken in the controller design and more sophisticated analysis of the system dynamics needs to be performed.

It is assumed that the warehouse is initially empty, i.e. $y(0) = 0$, and the first order is placed at $k = 0$, i.e. $u(k) = 0$ for $k < 0$. Due to delay, the first order is realized in period n_1, and $y(k) = 0$ for $k \leq n_1$. Taking into account the initial conditions, the stock level for any $k \geq 0$ may be calculated from the following equation

$$\begin{aligned} y(k) &= \sum_{p=1}^{m} \sum_{j=0}^{k-1} \lambda_p \varepsilon_p(j) u(j - n_p) + \sum_{p=1}^{m} \sum_{j=0}^{k-1} \lambda_p [1 - \varepsilon(j - n_p^R)] u(j - n_p - n_p^R) - \sum_{j=0}^{k-1} h(j) \\ &= \sum_{p=1}^{m} \sum_{j=k-n_p^R}^{k-1} \lambda_p \varepsilon_p(j) u(j - n_p) + \sum_{p=1}^{m} \sum_{j=0}^{k-1-n_p-n_p^R} \lambda_p u(j) - \sum_{j=0}^{k-1} h(j). \end{aligned} \tag{4}$$

3 Proposed Control Strategy

While formulating a control strategy for a dynamic system one typically concentrates on the evolution of the output variable, i.e. the quality measures (overshoots, oscillations, steady-state error, etc.) are evaluated with respect to y. In the considered application, due to the risk of triggering the bullwhip effect (increased order-to-demand variance ratio [14]), the quality of the generated control signal u is equally important. The established control signal should be smooth despite unknown demand variations and *a priori* unknown, variable failure rate, which poses a serious design challenge due to the presence of delay. In this paper, it is proposed to solve the considered control problem by applying a carefully chosen delay compensation mechanism. The proposed mechanism is illustrated in Fig. 3.

3.1 Delay Compensation Mechanism

In the considered system, in the linear region, the process of goods accumulation can be described by an integrator $G(z) = 1 / (z - 1)$. The proposed delay compensation mechanism consists of two structures. The first structure resembles the classical Smith predictor adapted for the analyzed case of multiple input-output channels with disparate delay $n_p + n_p^R$. The second structure uses the information about the actually admitted goods $\tilde{u}_p$.

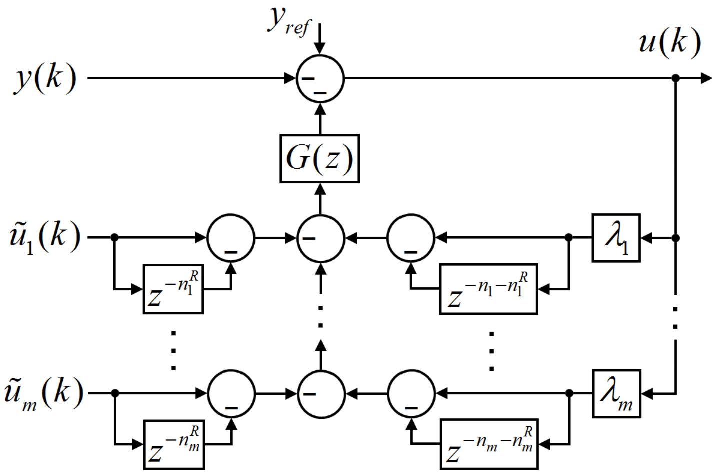


Fig. 3. Proposed delay compensation mechanism

3.2 Proposed Control Strategy

By referring to Fig. 3, the order quantity in period *k* is calculated from the following equation

$$u(k) = y_{ref} - y(k) - \sum_{p=1}^{m} \lambda_p \sum_{j=k-n_p-n_p^R}^{k-1} u(j) + \sum_{p=1}^{m} \sum_{j=k-n_p^R}^{k-1} \tilde{u}_p(j). \tag{5}$$

Alternatively, using (2), expression (5) can be represented as

$$u(k) = y_{ref} - y(k) - \sum_{p=1}^{m} \lambda_p \sum_{j=k-n_p-n_p^R}^{k-1} u(j) + \sum_{p=1}^{m} \sum_{j=k-n_p^R}^{k-1} \lambda_p \varepsilon_p(j) u_p(j-n_p). \tag{6}$$

3.3 Properties of Proposed Strategy

The properties of proposed inventory control strategy will be defined in a lemma and three theorems. The lemma and the first theorem show that the order quantities determined using the developed strategy are nonnegative and bounded, which is a crucial requirement for the practical implementation of any inventory management scheme. The second proposition specifies the warehouse capacity that needs to be provided to always accommodate the on-hand stock and incoming shipments. Finally, the third theorem indicates how to select the reference stock value in order to ensure maximum service level.

First, notice that since it is assumed that $u(k < 0) = 0$ and $y(0) = 0$, one has $u(0) = y_{ref}$. Afterwards, for $k \geq 1$, the control signal satisfies the relation given in the following lemma.

Lemma 1. *If policy* (6) *is applied to system* (3)*, then for any* $k \bullet 1$,

$$u(k) = h(k-1). \tag{7}$$

Proof. Substituting (4) into (6), one gets after algebraic manipulations

$$u(k) = y_{ref} - \sum_{p=1}^{m} \lambda_p \sum_{j=0}^{k-1} u(j) + \sum_{j=0}^{k-1} h(j). \tag{8}$$

Since $\sum_{p=1}^{m} \lambda_p = 1$, formula (8) can be further simplified in the following way

$$u(k) = y_{ref} - \sum_{j=0}^{k-1} u(j) + \sum_{j=0}^{k-1} h(j). \tag{9}$$

For $k = 1$, it follows immediately from (9) that $u(1) = y_{ref} - u(0) + h(0) = h(0)$, which shows that the lemma is indeed satisfied for $k = 1$. Let us assume that (7) is true for all integers up to some $l > 1$. Using this assumption, from (9), the order quantity generated in period $l + 1$ can be expressed as

$$\begin{aligned} u(l+1) &= y_{ref} - \sum_{j=0}^{l} u(j) + \sum_{j=0}^{l} h(j) = y_{ref} - u(0) - \sum_{j=1}^{l} u(j) + \sum_{j=0}^{l} h(j) \\ &= y_{ref} - y_{ref} - \sum_{j=1}^{l} h(j-1) + \sum_{j=0}^{l} h(j) = -\sum_{j=0}^{l-1} h(j) + \sum_{j=0}^{l} h(j) = h(l). \end{aligned} \tag{10}$$

Using the principle of the mathematical induction one may conclude that (7) actually holds true for arbitrary positive integer. This conclusion ends the proof. □

It follows from Lemma 1 that when the proposed strategy is applied, then the generated ordering signal is unaffected by the faults in the supply channels. Moreover, the variance of u,

$$\mathrm{var}(u) = \mathrm{var}(h) \leq (d_{\max} / 2)^2. \tag{11}$$

Consequently, since the order-to-demand variance ratio is smaller than or equal to 1, the bullwhip effect measured according to [14] is averted.

Theorem 2. *The order quantities generated by policy* (6) *applied to system* (3) *are always bounded, and for any* $k \bullet 0$ *the ordering signal satisfies the following inequalities*

$$0 \leq u(k) \leq \max\{y_{ref}, d_{\max}\}. \tag{12}$$

Proof. The initial order $u(0) = y_{ref}$, which means that the theorem is satisfied for $k = 0$. On the other hand, since $0 \le h(\cdot) \le d_{\max}$, from Lemma 1 one gets for any $k > 0$,

$$0 \le u(k) \le d_{\max}. \tag{13}$$

This conclusion ends the proof. □

Theorem 2 specifies the limits of control signal that can be expected in the proposed control system. In particular, it follows from (12) that no matter the demand and supply channel uncertainty, u never becomes negative. As a result, a feasible ordering signal is always ensured, even for the case of complete delivery failure $\varepsilon_p(k) = 0$. The next theorem shows how one should select the warehouse space at the distribution center to store all the incoming shipments.

Theorem 3. *If policy* (6) *is applied to system* (3), *then the stock never exceeds the reference level* y_{ref}.

Proof. Due to delay, the first shipments may reach the distribution center no sooner than in period $k = n_1$. Consequently, it follows from the assumed initial conditions, $y(0) = 0$ and $u(k) = 0$ for $k < 0$, that the warehouse is empty for any $k \bullet n_1$. Hence, it suffices to show that the proposition is satisfied for all $k \bullet n_1 + 1$. Using (4), the stock level in arbitrary period k may be expressed as

$$y(k) = \sum_{p=1}^{m} \lambda_p u(0) + \sum_{p=1}^{m} \sum_{j=1}^{k-1-n_p-n_p^R} \lambda_p u(j) + \sum_{p=1}^{m} \sum_{j=k-n_p^R}^{k-1} \lambda_p \varepsilon_p(j) u(j-n_p) - \sum_{j=0}^{k-1} h(j). \tag{14}$$

Since $\sum_{p=1}^{m} \lambda_p = 1$ and $u(0) = y_{ref}$,

$$y(k) = y_{ref} + \sum_{p=1}^{m} \sum_{j=1}^{k-1-n_p-n_p^R} \lambda_p u(j) + \sum_{p=1}^{m} \sum_{j=k-n_p-n_p^R}^{k-1-n_p} \lambda_p \varepsilon_p(j+n_p) u(j) - \sum_{j=0}^{k-1} h(j). \tag{15}$$

Then, using Lemma 1, one gets

$$\begin{aligned} y(k) &= y_{ref} + \sum_{p=1}^{m} \sum_{j=1}^{k-1-n_p-n_p^R} \lambda_p h(j-1) + \sum_{p=1}^{m} \sum_{j=k-n_p-n_p^R}^{k-1-n_p} \lambda_p \varepsilon_p(j+n_p) h(j-1) - \sum_{j=0}^{k-1} h(j) \\ &= y_{ref} + \sum_{p=1}^{m} \sum_{j=k-1-n_p-n_p^R}^{k-2-n_p} \lambda_p \varepsilon_p(j+n_p+1) h(j) - \sum_{p=1}^{m} \sum_{j=k-1-n_p-n_p^R}^{k-1} \lambda_p h(j). \end{aligned} \tag{16}$$

Finally, since $\varepsilon_p(\cdot) \le 1$ and $h(\cdot) \ge 0$,

$$y(k) \le y_{ref} - \sum_{p=1}^{m} \sum_{j=k-1-n_p}^{k-1} \lambda_p h(j) \le y_{ref}. \tag{17}$$

This conclusion finishes the proof. □

Theorem 3 specifies the upper limit of on-hand stock ever accumulated at the distribution center. Below, another important theorem is formulated. The theorem indicates how the value of y_{ref} should be selected to ensure that $y(k) > 0$. Consequently, one arrives at the state in which – after serving the imposed demand – the stock level is positive, and full demand satisfaction is obtained. Therefore, in addition to maximizing the profits from realized sales the company gains in the market credibility.

Theorem 4. *If policy* (6) *is applied to system* (3) *with* $0 < \underline{\ }_p \leq {}_p(k) \leq 1$, *and the reference stock level satisfies the following inequality*

$$y_{ref} > d_{\max} \sum_{p=1}^{m} \lambda_p [1 + n_p + (1 - \underline{\varepsilon}_p) n_p^R], \tag{18}$$

then for any $k \geq n_m + n_m^R + 1$ *the on-hand stock is strictly positive.*

Proof. Since ${}_p(\cdot) \geq \underline{\ }_p$ and $h(\cdot) \leq d_{\max}$, it follows from (16) that

$$\begin{aligned} y(k) &\geq y_{ref} + d_{\max} \sum_{p=1}^{m} \sum_{j=k-1-n_p-n_p^R}^{k-2-n_p} \lambda_p \underline{\varepsilon}_p - d_{\max} \sum_{p=1}^{m} \sum_{j=k-1-n_p-n_p^R}^{k-1} \lambda_p \\ &= y_{ref} - [d_{\max} \sum_{p=1}^{m} \lambda_p (1 + n_p + n_p^R) - d_{\max} \sum_{p=1}^{m} \lambda_p \underline{\varepsilon}_p n_p^R]. \end{aligned} \tag{19}$$

Consequently, using assumption (18), one gets for any $k \geq n_m + n_m^R + 1$, $y(k) > 0$. This conclusion ends the proof. □

4 Numerical Example

The properties of inventory control policy proposed in this paper are verified in simulations performed for the system described in Section 2. The review period is set equal to 1 day. It is assumed that three supply options ($m = 3$) are used for stock replenishment. The parameters characterizing each option are grouped in Table 1.

Table 1. System parameters

Option p	Lead time n_p	Repair time n_p^R	Uncer. bound $\underline{\varepsilon}_p$	Split coef. λ_p
1	5	5	0.77	0.32
2	6	6	0.85	0.28
3	7	7	0.82	0.40

With the maximum daily demand $d_{\max} = 90$ units the reference stock level is selected according to the guidelines of Theorem 3 so that full demand satisfaction is obtained. Consequently, $y_{ref} = 740 > 738$ units is chosen. The actual demand follows the pattern depicted in Fig. 4 (dotted curve), which reflects sudden changes in the market trend. The uncertainty associated with each channel is assumed uniformly distributed in the interval [ε_p, 1].

The ordering decisions taken by the controller are illustrated in Fig. 4 (continuous line), and the on-hand stock in Fig. 5. It is clear from Fig. 4 that the proposed controller closely follows the demand pattern without overshoots or oscillations. On the other hand, it follows from Fig. 5 that the stock level remains finite not exceeding y_{ref}, and after the initial phase, it does not fall to zero. Consequently, since $y(k) > 0$, the imposed demand is entirely satisfied from the readily available resources and maximum service level is achieved.

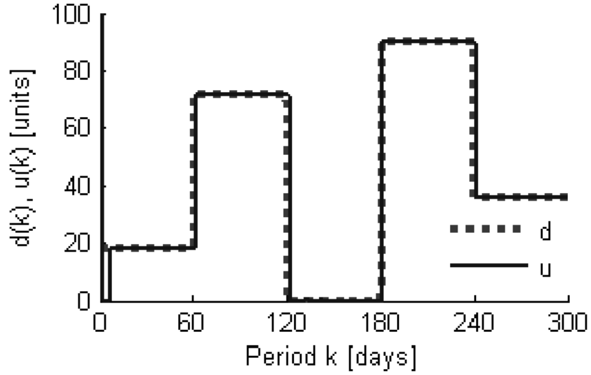


Fig. 4. Demand (dotted line) and orders (continuous line)

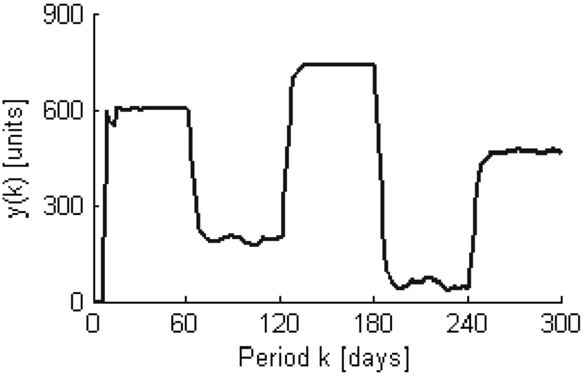


Fig. 5. On-hand stock level

5 Conclusions

In this paper, the problem of establishing an efficient stock replenishment strategy for production-inventory systems with faulty supply channels was addressed. The setting was modeled as an uncertain, parameter-varying system with multiple input-output delays. In order to counteract the negative influence of delay, a new compensation mechanism was designed. The proposed control strategy was demonstrated to provide fast reaction to varying market conditions with order-to-demand variance ratio not exceeding one. Thus, the bullwhip effect is averted despite uncertainty in goods delivery process. The limits of control signal (orders) and output variable (on-hand stock) were specified, and conditions for achieving full satisfaction of the *a priori* unknown demand were formulated and strictly proved.

Acknowledgements. This work has been performed in the framework of a project "Design and validation of control algorithms in networked dynamical systems" financed by the National Science Centre of Poland – decision number DEC-2012/05/D/ST6/03030. Support provided by the Foundation for Polish Science under Prof. A. Bartoszewicz's "Mistrz" grant is also acknowledged.

References

1. Ortega, M., Lin, L.: Control theory applications to the production-inventory problem: a review. Int. J. Prod. Res. 42, 2303–2322 (2004)
2. Sarimveis, H., Patrinos, P., Tarantilis, C.D., Kiranoudis, C.T.: Dynamic modeling and control of supply chain systems: a review. Computers Oper. Res. 35, 3530–3561 (2008)
3. Minner, S.: Multiple-supplier inventory models in supply chain management: a review. Int. J. Prod. Econ. 81-82, 265–279 (2003)
4. Dullaert, W., Vernimmen, B., Raa, B., Witlox, F.: A hybrid approach to designing inbound-resupply strategies. IEEE Intel. Syst. 20, 31–35 (2005)
5. Blanchini, F., Pesenti, R., Rinaldi, F., Ukovich, W.: Feedback control of production-distribution systems with unknown demand and delays. IEEE Trans. Robotics Autom. 16, 313–317 (2000)
6. Hoberg, K., Bradley, J.R., Thonemann, U.W.: Analyzing the effect of the inventory policy on order and inventory variability with linear control theory. Eur. J. Oper. Res. 176, 1620–1642 (2007)
7. Boccadoro, M., Martinelli, F., Valigi, P.: Supply chain management by H-Infinity control. IEEE Trans. Autom. Sci. Eng. 5, 703–707 (2008)
8. Ignaciuk, P., Bartoszewicz, A.: LQ optimal sliding mode supply policy for periodic review inventory systems. IEEE Trans. Autom. Control 55, 269–274 (2010)
9. Ignaciuk, P., Bartoszewicz, A.: Linear-quadratic optimal control strategy for periodic-review inventory systems. Automatica 46, 1982–1993 (2010)
10. Ignaciuk, P., Bartoszewicz, A.: LQ optimal and reaching law based sliding modes for inventory management systems. Int. J. Syst. Sci. 43, 105–116 (2012)
11. Ignaciuk, P.: Dead-time compensation in continuous-review perishable inventory systems with multiple supply alternatives. J. Proc. Control 22, 915–924 (2012)
12. Gu, K., Kharitonov, V.L., Chen, J.: Stability of time-delay systems. Birkhäuser, Boston (2003)
13. Normey-Rico, J.E., Camacho, E.F.: Dead-time compensators: a survey. Control Eng. Prac. 16, 407–428 (2008)
14. Chen, C., Drezner, Z., Ryan, J.K., Simchi-Levi, D.: Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, lead times, and information. Man Sci. 46, 436–443 (2000)

Application of Hierarchical Systems Technology in Conceptual Design of Biomechatronic System

Kanstantsin Miatliuk[1] and Yoon Hyuk Kim[2]

[1] Bialystok University of Technology, Dept. of Automation and Robotics, Wiejska 45C,
15-351 Bialystok, Poland
k.miatliuk@pb.edu.pl
[2] Kyung Hee University, School of Engineering, Yongin, 446-701, Korea
yoonhkim@khu.ac.kr

Abstract. Application of cybernetic technology of conceptual design of biomechatronic surgical robotics system (SRS) originated by systems theory is suggested in the paper. Traditional models of artificial intelligence and mathematics do not allow describing biomechatronic systems being designed on all its levels in one common formal basis, i.e. they do not give connected descriptions of the systems structure, the system as dynamic unit in its environment and the environment construction. To avoid this drawback, the technology of hierarchical and dynamic systems was chosen as a theoretical means for conceptual design and control tasks performing. Furthermore, in comparison with traditional methods the proposed technology gives the description of connected electronic, mechanical, biological, human-machine and computer subsystems of biomechatronic object in common formal basis. Theoretical basis of the coordination technology is briefly considered first. The example of the technology realization in conceptual design of SRS is presented after that in the paper.

Keywords: design, hierarchical system, mechatronics, surgical robot.

1 Introduction

In design process of biomechatronic systems we deal with objects which contain connected mechanical, electromechanical, biological, electronic, computer and human-computer subsystems. Various methods and models which used for each system coordination (design&control) can't describe all the subsystems and the mechanism of their interaction in the structure of higher level in common theoretical basis, and at the same time describe the system as a unit in its environment and environment system as well. It is important to define the common conceptual model which will describe all the above mentioned systems of biomechatronic object being designed in common formal basis. This task is topical for the systems of computer aided design (CAD) and processes of conceptual design in particular [1].

Conceptual model being created before the phase of detailed design [1] must also be coordinated with numerical and geometrical systems, i. e. the traditional forms of information representation in mechatronics. The theoretical basis of design process in

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_8, © Springer International Publishing Switzerland 2014

agreement with these requirements must be a hierarchical construction connecting any level unit with its lower and higher levels.

Mathematical and cybernetic theories based on the set theory are incoherent with the above requirements since the set theory describes one-level outlook. So, the coordination technology of Hierarchical System by Mesarovich and Takahara [2] with its standard block *aed* (ancient Greek word) by Novikava and Miatliuk [3-8] was chosen in the work as theoretical basis for the performing of conceptual design task of biomechtronic systems.

Nowadays, there are a lot of definitions of conceptual design [9,10]. The most precise from our point of view are "Conceptual design or what some call 'ideation' defines the general description of the product" given by Paul Brown, director of NX marketing for Siemens PLM Software and "the early part of any design process, which can occur at any point in the product development cycle" given by Bob McNeel [9]. M.J. French defines the conceptual design phase of the design process as the phase where the statement of the problem and generation of broad solution to it in the form of schemes is performed [10]. We recognize Conceptual Design as the process of creation of the systemic model of the object being designed at the early stage of its life cycle which is before the detailed design stage of object's mathematical model creation and numeric calculations realization.

In frames of the proposed coordination technology to create conceptual model for the design of biomechatronic object, i.e. a Surgical Robot System (SRS) in our case, means to define a SR system structure; its dynamic representation as the unit in its environment; SRS environment, its process and SRS-Environment interactions; SRS coordinator, its design & control processes; processes executed by SRS subsystems. Formal basis of conceptual design is given first in the paper. The exemplary conceptual model of Surgical Robot System is presented below after that.

2 Formal Model for Conceptual Design

Aed model S^{ℓ} considered below unites the codes of twolevel system [1] and general systems theory [11] by Mesarovic and Takahara, number code L^{s}, geometry and cybernetics methods; dynamic systems $(\overline{\rho},\overline{\varphi})$ are the main means of the description of the named codes. *Aed* is a standard element of hierarchical systems [3-8], which realizes the general laws of systems organization on each level and the inter-level connections. *Aed* S^{ℓ} contains ω^{ℓ} and σ^{ℓ} models connected by coordinator S_0^{ℓ}

$$S^{\ell} \leftrightarrow \{\,\omega, S_0\,, \sigma\}^{\ell} \tag{1}$$

where ω^{ℓ} is a dynamic representation of any level ℓ system in its environment, σ^{ℓ} is a system structure, S_0^{ℓ} is coordinator. Diagram of *aed* S^{ℓ} is presented in Fig.1.

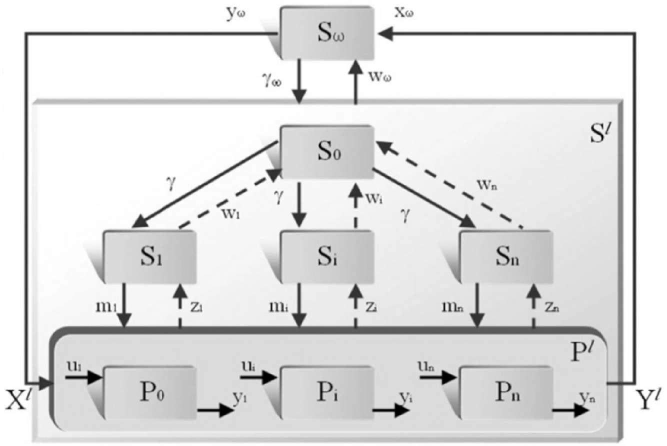


Fig. 1. Structure diagram of *aed* – standard block of hierarchical system. S_0 is the coordinator, S_ω is the environment, S_i are subsystems, P_i are subprocesses, P^l is the process of level ℓ, X^l and Y^l are the input and output of system S^l; m_i, z_i, γ, w_i, u_i, y_i are interactions.

Aggregated dynamic representations ω^ℓ of all *aed* connected elements, i.e. object ${}_oS^\ell$, processes ${}_oP^\ell$, ${}_\omega P^\ell$ and environment ${}_\omega S^\ell$ are presented in $(\overline{\rho},\overline{\varphi})^\ell$ form

$$\begin{aligned} \overline{\rho}^{\ell} &= \{\rho_t : C_t \times X_t \to Y_t \;\&\; t \in T\}^{\ell} \\ \overline{\varphi}^{\ell} &= \{\varphi_{tt'} : C_t \times X_{tt'} \to C_{t'} \;\&\; t,t' \in T \;\&\; t' > t\}^{\ell} \end{aligned} \tag{2}$$

where C^ℓ is state, X^ℓ is input, Y^ℓ is output, T^ℓ is the time of level ℓ, $\overline{\rho}^\ell$ and $\overline{\varphi}^\ell$ are reactions and state transition functions respectively. Object ${}_oS^\ell$, processes ${}_oP^\ell$, ${}_\omega P^\ell$ and environment ${}_\omega S^\ell$ are connected by their states, inputs and outputs [4-6].

Model of system structure is defined as follows

$$\sigma^\ell = \{ S_0^\ell ,\{ \overline{\omega}^{\ell-1} , {}_\sigma U^\ell \}\} = \{ S_0^\ell , \tilde{\sigma}^\ell \} \tag{3}$$

where S_0^ℓ is coordinator, $\overline{\omega}^{\ell-1}$ are aggregated dynamic models of subsystems $\overline{S}^{\ell-1} = \{S_i^{\ell-1} : i \in I^\ell\}$, ${}_\sigma U^\ell$ are structural connections. $\tilde{\sigma}^\ell$ is the connection of $\overline{\omega}^{\ell-1}$ systems and their interactions ${}_\sigma U^\ell$ coordinated with external ones ${}_\omega U^\ell$.

Coordinator S_0^ℓ is the main element of hierarchical systems which realizes the processes of systems design and control [3-8]. It is defined according to *aed* presentation of equation (1) in the following form

$$S_0^\ell = \{\, \omega_0^\ell, S_{00}^\ell, \sigma_0^\ell \,\} \tag{4}$$

where ω_0^ℓ is aggregated dynamic realization of S_0^ℓ, σ_0^ℓ is the structure of S_0^ℓ, S_{00}^ℓ is coordinator control element. S_0^ℓ is defined recursively. Coordinator S_0^ℓ constructs its aggregated dynamic realization ω_0^ℓ and structure σ_0^ℓ by itself. S_0^ℓ performs the design and control tasks on its choice, learning and self-organization strata [3-6].

All metric characteristics μ of systems being coordinated (designed & controlled) and the most significant geometry signs are determined in the frames of *aed* informational basis in codes of numeric positional system L^S [4-6]. Structures have two basic characteristics: ξ^ℓ (connection defect) and δ^ℓ (constructive dimension); μ^ℓ, ξ^ℓ and δ^ℓ are connected and described in positional code of L^S system [4-6]. For instance, constructive dimension $\delta^\ell \in \Delta^\ell$ of system S^ℓ is presented in L^S code as follows

$$\begin{gathered} \tilde{\delta}^\ell = (n_3 \ldots n_0)_\delta, \tilde{\delta}^\ell \in \{\delta_\sigma^\ell, \delta_\omega^\ell\} \\ (n_i)_\delta = (n_{3-i})_\xi\,,\ (n_i)_\delta \in N\,,\ i{=}0{,}1{,}2{,}3 \end{gathered} \tag{5}$$

where δ_ω^ℓ and δ_σ^ℓ are constructive dimension of σ^ℓ and ω^ℓ respectively. This representation of geometrical information allows execution of all operations with geometric images of mechatronic objects on computer as operations with numeric codes. *Aed* technology briefly described above presents theoretical basis for surgical robot system (SRS) conceptual design and control.

3 Conceptual Model of SR System

As an example of biomechatronic object description, the formal model of SR system is presented below. Recently SR systems such as commercial ROBODOC system (Integrated Surgical Systems, CA, USA) are widely used in TKA (Total Knee Arthoplasty) surgery. In the paper we focused on laboratory-level SRS with industrial robot and navigation system for TKA (Fig.2) developed in BioMech Lab., School of Engineering, Kyung Hee University (KHU), South Korea [12]. Conceptual systemic model of the SR system is presented in *aed* form

$$S^\ell \leftrightarrow \{\omega, S_0, \sigma\}^\ell \tag{6}$$

where ω^ℓ is an aggregated dynamic representation of SRS S^ℓ, σ^ℓ is the system structure, S_0^ℓ is SRS coordinator (design & control system), ℓ is the index of level.

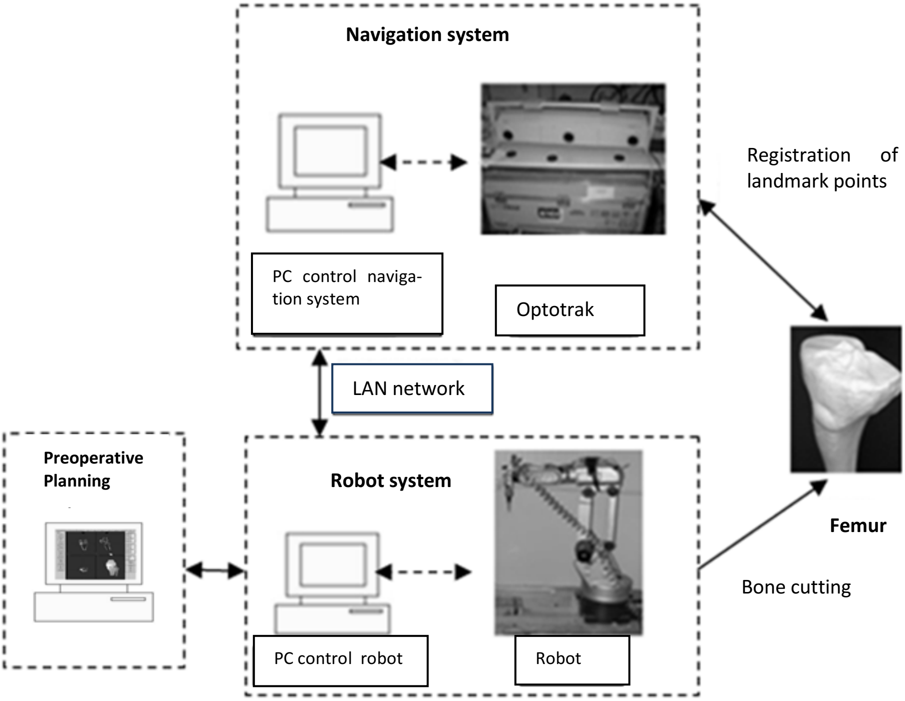


Fig. 2. Scheme of laboratory SRS for TKA

SR system structure σ^{ℓ} contains the set of sub-systems $\overline{\omega}^{\ell-1}$ and their structural connections ${}_{\sigma}U^{\ell}$. Thus, according to model (3) the SRS subsystems $\overline{\omega}^{\ell-1}$ are

$\omega_1^{\ell-1}$: robot system (RS); $\omega_2^{\ell-1}$: pre-operative planning system (PP);

$\omega_3^{\ell-1}$: navigation system (NS); $\omega_4^{\ell-1}$: computer control system (CCS).

In their turn, each subsystem has its own structural elements – lower level $\ell-2$ subsystems. For the robot mechatronic subsystem (RS) $\omega_1^{\ell-1}$ they are manipulator $\omega_{11}^{\ell-2}$ (mechanical), servomotors $\omega_{12}^{\ell-2}$ (electromechanical), cutting machine $\omega_{13}^{\ell-2}$ (pneumatic) and own control system $\omega_{14}^{\ell-2}$ (computer). CCS contains control units of robot, pre-operative planning and navigation subsystems of SRS and communication program developed to integrate the subsystems. CCS is a part of SRS coordinator which performs its control functions. All the subsystems are connected by their structural connections ${}_{\sigma}U^{\ell-2}$. For instance, cutting machine $\omega_{13}^{\ell-2}$ and manipulator $\omega_{11}^{\ell-2}$ are connected by ending effector ${}_{\sigma}U_{13}^{\ell-2}$ of robot. By analogy, the higher level subsystems $\overline{\omega}^{\ell-1}$ are connected by their common parts, i.e. structural connections ${}_{\sigma}U^{\ell-1}$ that are the elements of lower levels. Navigation system $\omega_3^{\ell-1}$ and robot

$\omega_1^{\ell-1}$ are connected by their common element – communication program of CCS ${}_{\sigma}U_{13}^{\ell-1} = \omega_{14}^{\ell-2} = \omega_{34}^{\ell-2}$, where $\omega_{14}^{\ell-2}$ is dynamic representation of the control program being the subsystem of robot $\omega_1^{\ell-1}$, and $\omega_{34}^{\ell-2}$ the one of the program being the subsystem of the NS $\omega_3^{\ell-1}$.

Aggregated dynamic realizations $\overline{\omega}^{\ell-1}$, i.e. dynamic models ${}_i(\overline{\rho},\overline{\varphi})^{\ell-1}$ of SRS subsystems are formed after the definition of their inputs-outputs concerning each concrete subprocess they execute. Thus, for the navigation system $\omega_3^{\ell-1}$, concerning its registration process of bone and robot the input $X_3^{\ell-1}$ is optic signal received by Optotrak system and the output $Y_3^{\ell-1}$ is the robot instrument and real bone coordinates collected in control PC. State $C_3^{\ell-1}$ in this case is the stage of registration process completeness. This process is of the informational nature. As for the robot subsystem $\omega_1^{\ell-1}$, concerning its mechanical process of instrument motion the realization $\omega_{11}^{\ell-1}$ can be presented in form (2) as equations [13] of inverse kinematics:

$$\dot{q} = J^+(q)\dot{x} \tag{7}$$

which connects robot joints velocities $\dot{q}$ as the output $Y_1^{\ell-1}$ with velocities $\dot{x}$ of robot instrument as input $X_1^{\ell-1}$, where J^+ is the pseudo inverse of Jacobian matrix.

Environment of the SRS has its own structure and contains the following subsystems: ω_1^{ℓ} bone (biological); ω_2^{ℓ} : surgeon, which communicates with SRS via video information, registration and cutting motion planning system (human-computer system); ω_3^{ℓ} other biomechatronic systems being in interaction with SRS (e.g. computer tomography CT system, which supplies PP subsystem with bone images); ω_4^{ℓ} implant producing system; ω_5^{ℓ} : higher level coordinator (design system).

Thus the surgery-SRS interactions (by obtaining video image of bone, checking a cutting path & robot motion planning), and robot cutting instrument-bone interactions are control and executive interactions of SRS and its environment respectively. The immediate input X^{ℓ} for the SR system (which is the output ${}_{\omega}Y^{\ell} = X^{\ell}$ of the environment of SR system) are control actions produced by surgeon and the inputs generated by other environment systems, e.g. inserted implant presented by ω_4^{ℓ}. The output Y^{ℓ} of the SR system is the cutting bone. The output of the SR system S^{ℓ} is at the same time the input ${}_{\omega}X^{\ell} = Y^{\ell}$ of the environment ${}_{\omega}S^{\ell}$. The states C_i^{ℓ} of SR system S^{ℓ} are the inputs of TKA surgical processes. Dynamic representation ω^{ℓ} of SRS and its subsystems are constructed in form of (2) by the inputs, states and outputs mentioned above. The dynamic representation $(\overline{\rho},\overline{\varphi})$ can be given at the stage of detailed design in form as follows

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx\,. \end{aligned} \tag{8}$$

First state equation in (8) corresponds to the state transition function $\overline{\varphi}$ in (2) and the output equation corresponds to the reaction $\overline{\rho}$. Vectors *x, y, u* and matrices *A, B, C* must be predefined. In the case of robot end effector motion the elements of states vector $x=[x_1\ x_2\ x_3]^T$ are the displacement x_1, velocity x_2 and acceleration x_3.

SRS TKA process P^{ℓ} is a part of higher-level process $P^{\ell+1}$ in environment ${}_{\omega}S^{\ell}$. Environment is a biomechatronic system of higher level which includes general design&control system, surgeon, person being operated, other biomechatronic systems being in interaction with SRS (CT system and implant production system).

This process contains the following sub-processes of level ℓ : (P_1) virtual model (VM) of bone and implant graphic image forming (by PP system); (P_2) real bone image (RBI) and robot instrument location registration (by NS); (P_3) VM matching to RBI model; (P_4) cutting path forming in PP and transforming to coordinate of RS; (P_5) allocation of robot instrument in the registered points by surgeon using control haptic device, and joints angels control law definition by solving inverse kinematics problem; (P_6) sending control signal to AS2 robot system; (P_7) AS2 robot motion and end effector (pneumatic instrument) displacement; (P_8) bone cutting and implant inserting. All the subprocesses are formally described according to (2).

P_7 is realized by electromechanical subsystems (Samsung AS2 robot servomotors) of general biomechatronics system (SR system), P_1 - P_4 realized by video-information and computer subsystems, and P_8 by pneumatic and mechanical ones. The general process is composed by subprocesses $\overline{P}^{\ell}$, executed by the general biomechatronic system, which includes the SR system S^{ℓ} and its environment ${}_{\omega}S^{\ell}$.

So, all the subsystems of general biomechatronic system, i.e. mechanical (manipulator $S_{11}^{\ell-2}$, haptic system), electromechanical (servomotor $S_{12}^{\ell-2}$), pneumatic (cutting instrument $S_{13}^{\ell-2}$), computer-electronic (navigation $S_3^{\ell-1}$ and control system $S_4^{\ell-1}$), human-computer (surgeon and PP $S_2^{\ell-1}$ system) have their aggregated dynamic ω^{ℓ} and structural σ^{ℓ} descriptions. All the connected descriptions of the subsystems $\overline{S}^{\ell}$ and processes $\overline{P}^{\ell}$ are presented in the informational resources (data bases) of the coordinator which realizes the design and control processes, connecting in this way structure σ^{ℓ} and dynamic realization ω^{ℓ} of the SRS being coordinated.

Coordinator S_0^ℓ is formally described by (4) and realized in the form of human-computer design&control system of the SRS, which maintains its functional modes by surgeon and control system and realizes the design process by higher level computer aided design (CAD) system. All metrical characteristics of SR subsystems necessary in design process are presented in the form of numeric positional systems L^s [4-6].

4 Realization of Coordination Technology

Coordinator S_0^ℓ tasks in control process P_0^ℓ of SRS are performed by human-computer system, i.e. surgeon in communication with servers and robot. The main control tasks are indicated in GUI window of control program developed in BioMech Lab., KHU, South Korea. The control processes are 1) bone fixation and its position checking ${}_1P_0^\ell$, 2) registration of robot position ${}_2P_0^\ell$, 3) registration of real-virtual bone relations ${}_3P_0^\ell$, 4) cutting path calculation and a bone cutting ${}_4P_0^\ell$.

Formally, control functions are described as coordinator S_0^ℓ functions in $(\overline{\rho},\overline{\varphi})_0^\ell$ form (2) after definition of states C_{0t}^ℓ, inputs X_{0t}^ℓ and outputs Y_{0t}^ℓ of SRS control subsystems. According to the scheme in Fig.1, the inputs and outputs of coordinator are defined on the sets of its coordination signals G^ℓ and feedbacks W^ℓ as follows:

$$X_0^\ell = \{ G^{\ell+1}, W^\ell \}, \qquad Y_0^\ell = \{ G^\ell, W^{\ell+1} \} \tag{9}$$

where G^ℓ are coordination signals for systems $\overline{S}^{\ell-1}$, W^ℓ is feedback from $\overline{S}^{\ell-1}$, $W^{\ell+1}$ is feedback from S_0^ℓ to coordinator $S_0^{\ell+1}$ of higher level, $G^{\ell+1}$ are coordination signals from $S_0^{\ell+1}$ to S_0^ℓ [4-6].

Concerning Multi Motion Control (MMC) system [14] of robot the input is the preplanning cutting trajectory of robot's instrument and the output is the electronic control signals (joints angular values) which run to manipulator (servomotors) to realize predefined motion of its joints after inverse kinematic problem solving. State of general control system is the state of control process in the current moment of time.

For MMC control process (Fig. 3) the inputs X_{0t}^ℓ are G^ℓ which are coordinates of cutting path *[x(t), y(t), z(t)]* and feedback signals $W^{\ell-1}$ from servomotors which bring the actual values of joints current positions. Outputs Y_{0t}^ℓ are $G^{\ell-1}$ electronic signals which define manipulator joint angular values θ_i (new position), *i=6*, and W^ℓ is feedback to PP system, i.e. position of robot and cutting instrument, registered by Optotrak. State C_{0t}^ℓ is the actual position of the manipulator joints and cutting instrument at *t* moment of time.

Concerning positioning and bone cutting process controlled by MMC system the input of AS3 Samsung robot $X_1^{\ell-1}$ are input loads of servomotors. Its output $Y_1^{\ell-1}$ is positioning&cutting operation which is the input of environment ${}_{\omega}S^{\ell}$ element ω_1^{ℓ}, i.e. bone being operated. This cutting process is executed by surgeon using GUI program activating control processes ${}_1P_0^{\ell-1}$ of MMC system. Control systems (${}_iS_0^{\ell-1}$ coordinators) of SRS sub-systems, i.e. NS, CCS, PP system, robot system, shown in Fig. 3 create the general SRS coordinator.

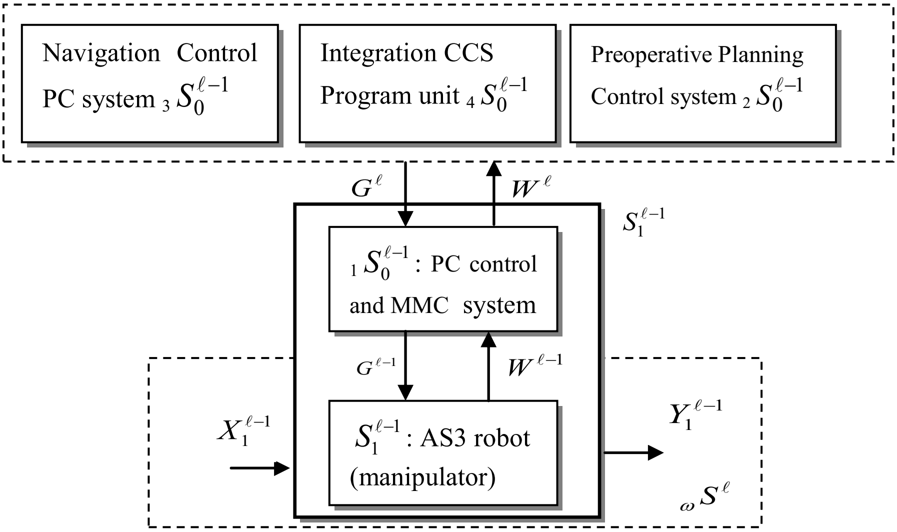

Navigation Control PC system ${}_3S_0^{\ell-1}$
Integration CCS Program unit ${}_4S_0^{\ell-1}$
Preoperative Planning Control system ${}_2S_0^{\ell-1}$
G^{ℓ}
W^{ℓ}
$S_1^{\ell-1}$
${}_1S_0^{\ell-1}$: PC control and MMC system
$G^{\ell-1}$
$W^{\ell-1}$
$X_1^{\ell-1}$
$S_1^{\ell-1}$: AS3 robot (manipulator)
$Y_1^{\ell-1}$
${}_{\omega}S^{\ell}$

Fig. 3. Interaction of SRS sub-systems in control process of manipulator realizing cutting path

5 Conclusions

The conceptual model of SRS biomechatronic system presented in the theoretical basis of hierarchical systems for the design and control technology realization is briefly presented in the paper. In comparison with traditional methods of mathematics and artificial intelligence the proposed SRS formal model contains connected descriptions of the coordinated (designed&controlled) system structure, its aggregated dynamic representation as unit in its environment and the environment model. All the descriptions are connected by the coordinator which performs the design and control tasks on its strata. The model presented is coordinated with traditional systems of information presentation in biomechatronics: numeric, graphic and natural language forms [7]. *Aed* technology is also coordinated with general requirements of design and control systems [3-7], considers the elements of SRS conceptual model as well as connected SRS subsystems of different nature (mechanical, electromechanical, electronic, human-computer, biological) in common *aed* theoretical basis. It brings new possibilities in creating of a formal language for conceptual design.

Besides, the proposed technology makes the transfer from conceptual to detailed design step in the design process very easy. Because of the given conceptual model contains as its elements the dynamic systems $(\overline{\rho},\overline{\varphi})^{\ell}$ which are the generalization of mathematical models (DE, automata, algebra systems [11]) the transfer algorithm to the next design step is only the concretization of these models. It allows to rise the efficiency of the whole design process of engineering systems and biomechatronic systems in particular. Given technology was also applied for design and control of other engineering objects [8,15], in biomechanics [16] and mechatronics [17].

References

1. Moulianitis, V., Aspragathos, N., Dentsoras, A.: A model for concept evaluation in design – an application to mechatronic design of robot gripper. Mechatronics 14, 599–622 (2004)
2. Mesarovic, M., Macko, D., Takahara, Y.: Theory of Hierarchical Multilevel Systems. Academic Press, New York (1970)
3. Novikava, S., Mialtiuk, K., Gancharova, S., Kaliada, W.: Aed Construction and Technology in Design. In: 7th IFAC LSS Symposium, pp. 379–381. Pergamon, London (1995)
4. Novikava, S., Mialtiuk, K., et al.: Aed Theory in Hierarchical Knowledge Networks. Studies in Informatics and Control 6(1), 75–85 (1997)
5. Novikava, S., Miatliuk, K., et al.: Mathematical and Cybernetical Means of Aed Theory. In: IEEE Int. Conf. on Systems, Man, and Cybernetics, vol. 4, pp. 2874–2879 (1996)
6. Mialtiuk, K.: Coordination Processes of Geometric Design of Hierarchical Multilevel Systems. Machine Constructing and Exploitation 11, 163–178 (2003) (in Polish)
7. Novikava, S., Mialtiuk, K.: Hierarchical System of Natural Grammars and Process of Innovations Exchange in Polylingual Fields. Kybernetes 36(5/6), 736–748 (2007)
8. Miatliuk, K.: Coordination Technology in Design of Biologically Inspired Robot. Machine Dynamics Problems 33(3), 70–78 (2009)
9. Hudspeth, M.: Conceptual Design in Product Data Management. Design Engineering Technology News, Desktop Engineering (February 1, 2008)
10. French, M.J.: Conceptual Design for Engineers, 3rd edn. Springer (1999)
11. Mesarovič, M., Takahara, Y.: Abstract systems theory. Springer (1990)
12. Kim, Y.H., Minh, H.L.: A Laboratory-Level Surgical Robot System for Minimal Invasive Surgery Total Knee Arthroplasty. J. Precision Eng. and Manufacturing 12(2), 237–242 (2011)
13. Siciliano, B.: Kinematic Control of Redundant Robot Manipulators: A Tutorial. Journal of Intelligent and Robotic Systems 3(3), 201–212 (1990)
14. Multi Motion Controllers MMC Manual, Rockwell Samsung Automation Inc., Korea
15. Miatliuk, K., Kim, Y.H., Kim, K.: Motion Control Based on the Coordination Method of Hierarchical Systems. J. of Vibroengineering 11(3), 523–529 (2009)
16. Miatliuk, K., Siemieniako, F., Kim, Y.H., Kim, K.: Human Motion Design in Hierarchical Space. In: 16th ICSS Conference on Systems Science, Wroclaw, pp. 496–503 (2007)
17. Miatliuk, K., Kim, Y.H., Kim, K., Siemieniako, F.: Use of Hierarchical System Technology in Mechatronic design. Mechatronics 20(2), 335–339 (2010)

Questions of Synthesis Systems Precision

Darja Gabriska*, Augustin Gese, and Lubos Ondriga

Slovakia, Slovak University of Technology in Bratislava,
Faculty of Materials Science and Technology in Trnava
darja.gabriska@gmail.com,
{augustin.gese,lubos.ondriga}@stuba.sk

Abstract. This article discusses the basic principles of synthesis of precision control systems and the main challenges that limit the dynamic accuracy of these systems.

Keywords: disturbance control, control precision, system sensitivity, drives, stability analysis.

1 Introduction

Requirements for maximum interference suppression, arising from the object and motor systems in stabilizing systems of the rotational speed drives with high accuracy is one of the most important roles in problem solution the movement accuracy of working mechanism.. In general, this is achieved by extending the control bandwidth and increasing astatism system, or by a combined control [1]. This article describes the basic principles of synthesis of control systems a of high-precision and fundamental problems that limit the dynamic accuracy of these systems.

The basic way to enlarge accuracy of the motion stabilization is to increase the control bandwidth. However, the extension of the control bandwidth is always restricted by the resonant characteristics of mechanical structures. Currently, this limitation is the most important. Another way to increase the accuracy of motion stabilization is designing a combined control, which is associated with direct or indirect identification of disturbance. This solution is not always effective in the case of low-level interference and unnecessary complication of the control system. In addition, other disorders that arise with complicated structures often completely suppress the expected positive effect.

2 Design of Dynamics Characteristics

Considering the control systems design principles and obtaining the required dynamic characteristics, the principle cascade control from the engineering point of view is the most convenient and effective. This principle assumes linearity of characteristics of

* Corresponding author.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_9, © Springer International Publishing Switzerland 2014

the of the control system elements. For the control systems with high accuracy, this principle only works for the first phase of the synthesis, while the second synthesis phase is absolutely necessary to consider control system elements such as saturation.

Increasing the range of astatism of outer loop of control improves static and dynamic accuracy of the rotational speed stabilization. Synthesis of the cascade control systems generally starts from the inner loop. In our case, we proceeded oppositely. We select the desired transfer function of the system. The frequency of intersection was determined considering the mechanical resonance frequency of the structure. Taking into account the structure of the control system, which has astatism of the second row with PID controller, the required transfer function W (s) has 3-1-2-4 shape [4, 5].

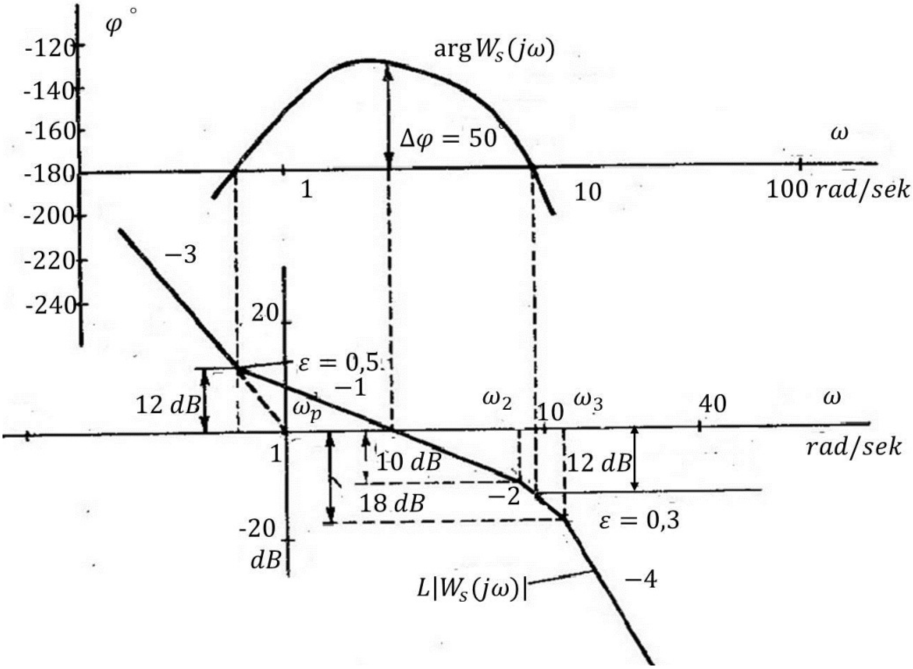


Fig. 1. Bode plot of the control system

The Fig.1 shows optimized Bode diagram of the required frequency characteristic of the system with frequency of the intersection $\omega_P = 1$. The above diagram corresponds to the transfer function

$$W_K(s) = \frac{2{,}56\left(\frac{s}{\omega_p}\right)^2 + 1{,}6\left(\frac{s}{\omega_p}\right) + 1}{\left(\frac{s}{\omega_p}\right)^3 \left(0{,}125\frac{s}{\omega_p} + 1\right)\left[\left(0{,}08\frac{s}{\omega_p}\right)^2 + 2 \cdot 0{,}3 \cdot 0{,}08 \cdot \frac{s}{\omega_p} + 1\right]} \quad (1)$$

Because it is technically possible to carry out high dynamic characteristics of the amplifier current (torque), limiting the control bandwidth depends only on the resonance properties of the drive (TR). The transfer characteristic of the PID controller for the system (1) has the form

$$W_{pid}(s) = \frac{\beta_\varphi\left[T_\varphi T_d (s)^2 + T_\varphi s + 1\right]}{s\left(T_f s + 1\right)} \tag{2}$$

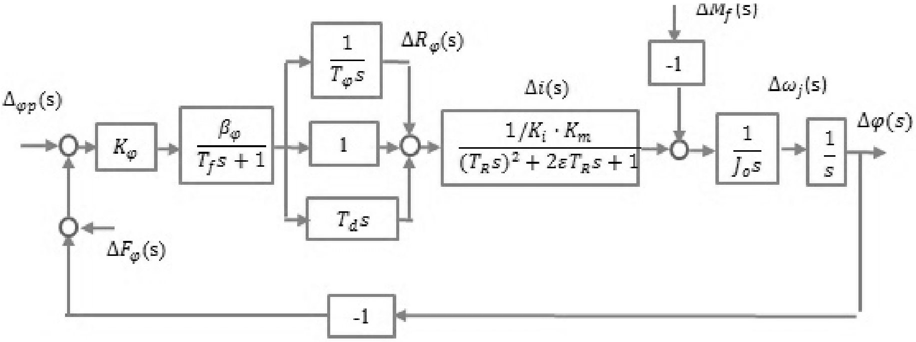


Fig. 2. Block diagram of the system

Block diagram of the system (fig.2) corresponds to the transfer function of an uncoupled position control loop

$$W_k(s) = \frac{K_\varphi \cdot K_m \cdot \beta_\varphi \cdot \left[T_\varphi \cdot T_d \cdot (s)^2 + T_\varphi \cdot s + 1\right]}{K_i \cdot J_o \cdot T_\varphi \cdot (s)^3 \cdot \left(T_f \cdot s + 1\right)\left[(T_R \cdot s)^2 + 2 \cdot \varepsilon \cdot T_R \cdot s + 1\right]} \tag{3}$$

By the comparison of the transfer function (1) and the modified one (3) we can write the necessary values to tune the system

$$T_R = \frac{0{,}08}{\omega_p};\ T_f = \frac{0{,}125}{\omega_p};\ T_\varphi \cdot T_d = \frac{2{,}56}{{\omega_p}^2}; T_\varphi = \frac{1{,}6}{\omega_p};\ {\omega_p}^2 = \frac{K_\varphi \cdot K_m \cdot \beta_\varphi}{1{,}6 \cdot J_o \cdot K_i} \tag{4}$$

In the known resonant characteristics of the driver (T_R) a is possible to determine the parameters of the position controller

$$\omega_p = \frac{0{,}08}{T_R};\ T_f = \frac{0{,}125}{\omega_p};\ T_\varphi = \frac{1{,}6}{\omega_p}; T_d = \frac{2{,}56}{T_\varphi \cdot {\omega_p}^2};\ \beta_\varphi = 1{,}6 \cdot \frac{K_i \cdot J_o}{K_\varphi \cdot K_m} \cdot {\omega_p}^2 \tag{5}$$

Analysis of the optimal value of the proportional component of the position controller to the minimum quantization noise [1] requires $\beta_\varphi = 1$. Then we can determine the optimal sensitivity of the phase measurement

$$K_\varphi = 1{,}6 \cdot \frac{K_i \cdot J_o \cdot {\omega_p}^2}{K_m} \tag{6}$$

or contents of the crossover frequency of the transfer function of the control systems in the known sensitivity coefficient measurement phase K_φ

$$\omega_p = \sqrt{\frac{K_\varphi \cdot K_m}{1{,}6 \cdot K_i \cdot J_o}} \tag{7}$$

An analysis of formulas (5,6,7) results in two possible algorithms of the selection of controller parameters.

First. The structure of the drive is given (known resonant characteristic of drive) and we have to design parameters of controller and interpolator, which are admissible. We have requirements for accuracy of stabilization, and we have to propose the construction of drive (depending on the desired resonant characteristics of the drive), parameters of the controller and interpolator, which we derive from the resonance frequency of the drive.

Second. In the case of drive without bearings [2] resonant frequency may no longer be a factor limits. Then, we can just limit. Then, we can just limit of the dynamic characteristics of the drive. Then, we can just limit dynamic options (Ti) current control circuit (torque) to calculate the base frequencyω_p. For the limitation criteria of the dynamic possibilities of current control circuit we have to consider the fact that the extension of the control circuit bandwidth increases the interfering component of current (torque). High frequency of interfering component of current only warms the drive. The source of the interfering component of current is quantization noise of interpolator, which can be considered a type of random "white noise".

Quantization-noise dispersion [1] is $D_k = N \cdot \Delta f = \frac{\Delta^2}{12}$, where Δ – quantization step of interpolator,

N – spectral noise density and Δf –bandwidth noise. For numeric type of interpolator, frequency band noise is determined by sampling rate, where T_o – sampling period.

$$f_o = \frac{1}{2\pi T_o}, \qquad N = \frac{\Delta^2}{12} \cdot \left(\frac{1}{2\pi T_o}\right)^2 \tag{8}$$

Then the dispersion component of the torque drive noise, quantization noise source the source of with spectral density N [4] is:

$$D_m(\omega) = N \cdot \int_0^\infty \left| \frac{(J \cdot j\omega)^2}{1 + W_k(j\omega)} \right|^2 d\omega \tag{9}$$

Where J – mechanical inertia drive.

To solve the function (9) we can by the solution of the integral [4]

$$I_n = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{G(j\omega)}{|A(j\omega)|^2} \tag{10}$$

Where

$$A(j\omega) = a_0(j\omega)^n + a_1(j\omega)^{n-1} + \cdots + a_n \tag{11}$$

$$(j\omega) = b_0(j\omega)^{2n-2} + b_1(j\omega)^{2n-4} + \cdots + b_{n-1} \tag{12}$$

Solution of integrals of type (11) to n = 7 as a table of coefficients of formula (10) can be found in the literature, for example in [6]. So, when we choose the quantization step Δ and sampling frequency then we can calculate the disturbance dispersion and thermal load on the motor. Subsequent work by limiting bandwidth circuit current control can limit the thermal load on the motor. An alternative way of circuit characteristics design is to determine the quantization step and the sampling period for the desired operating band of the current control circuit.

3 Compensation of Periodically Repeated Disturbance

We discuss the behavior of the automatic control system with compensation device periodically repeated disturbance. The general model of such a control system is shown in Figure 3

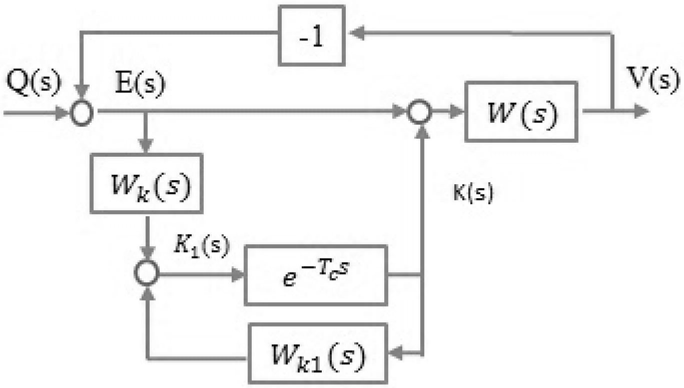


Fig. 3. Block diagram of control with compensation of periodic disturbance

The picture shows the following information:

Q(s) – set value, E(s) – control deviation, W_K (s)– filter transfer function in the error signal; W_{K1}(s) - transfer function of the filter compensation signal; W(s) – transfer function control system; K(s) – signal compensation; K_1(s) – signal at the compensation loop input; T_C – period of the compensation cycle.

Time delay in the signal memory is equal Tc and usually considerably exceeds the time of the transitional process in main control loop. It follows that the establishment of compensating device does not violate the stability of the loop. The presence the presence of multiple correction of compensation signal requires the analysis of the convergence process of compensation. [2].

Equation of the deviation of automatic control system (Fig.3) at periodically repeated failures disturbances

$$E(s) = \frac{E_p(s)}{1 - e^{sT_c}} = \Phi_\varepsilon(s)\frac{Q_p(s)}{1 - e^{sT_c}} - \Phi_\varepsilon(s)W_f(s)\frac{F_p(s)}{1 - e^{sT_c}} \tag{13}$$

where $E_p(s)$- periodic component of deviation, T_c- time of the cycle of deviation and disorders repeat, $\Phi_\varepsilon(s) = 1/[1 + W(s)]$ – transfer characteristics of the control system according to deviation, W(s)- transfer characteristic of decoupled control loop, $Q_p(s)$ – set point of the control system periodic component, $W_f(s)$- transfer function according to disorders, $F_p(s)$ – transfer function of the periodic component of failures.

Periodic repetition of the control and disorder operations allows the use of periodic component of the dynamic error for the formation of an additional motion program, which compensates the effect of interference regardless of their nature and the point of application in the system, where control of disorders has the form

$$\frac{E_k(s)}{1 - e^{sT_c}} = \frac{E_p(s)}{1 - e^{sT_c}} - \Phi(s)\frac{K_p(s)}{1 - e^{sT_c}} \tag{14}$$

From (14) follows that the total compensation of periodic components of disturbances requires the implementation of a correction circuit with a transmission characteristic $W_k(s) = 1/\Phi(s)$ to calculate the compensation program for the periodic component of the deviation $E_p(s)$. The exact implementation of the transfer function $W_k(s)$ would require a multiple differentiation in the high frequency (greater than the intersection of frequency), which leads to a considerable noise accretion. Furthermore, the calculated values of the compensation effect at high frequencies is of particular amplitude, which means that it is possible to limit the compensation ratio in the case of physical implementation and thus breach the requirements for compensation. Therefore there have to be a practical limit of the compensation of the frequency component of the disturbance, which slightly exceeds the frequency of intersection of the control circuit and by an approximate implementation $W_k(s)$ of these frequencies.

In real system of automatic control, besides the standard components, the control deviation contains also random ones. In the high precision control systems, these components are comparable based on the size. There is therefore a need to use static processing of the measurement results of the control deviation before calculating compensatory interference. The solution of this task can be done in two ways. The first one is to calculate estimation of the mathematical expectation of automatic the automatic based on the results of measurement of the control deviation of a sufficiently large number of cycles of the deviation [1]. In this case, the compensation will begin after longer period, what results in the fact that the method can be used in systems that operate in a one mode for a longer time. Furthermore, at a change in the nature of disorder effects after the implementation of compensation, the correction accuracy is lost.

The second way is to start compensation after the first cycle treatment of the disorder with subsequent correction of compensatory according to the results setting of previous compensatory action. Reducing the impact of signal random component of the control deviation is achieved by introducing scale (less than one) at the input and output of the compensation circuit, which works as follows.

The result of the first cycle of disorder change processing (scale is equal to the one) is stored in memory and at the processing of the second cycle is added to the current value of the original program and this way it creates value of the correction program is of the sum of the old value and a certain proportion of the value of the current system disorder. In this way for several cycles of disorder change processing stored in memory the additional action is accumulated, what allows compensating of the regular component of control disorder. Transfer characteristic of the deviation of the automatic control a cyclical change of disorders for the system (3) has the form

$$W_{od}(s) = -[W_{k1}(s) - W_k(s)\Phi(s)] \cdot e^{-sT_c} \tag{15}$$

Then the transfer function of a closed-loop of the system under the deviation is as follows:

$$\Phi_{k\varepsilon}(\mathrm{s}) = \frac{1}{1 + \left[1 + W_k(s)W_{sf}(s)\right] \cdot W(s)} \tag{16}$$

Where $W_{sf}(s)$ – transfer function of synchronous filter

$$W_{sf}(s) = \frac{e^{-sT_C}}{[1 - W_{k1}(s) \cdot e^{-sT_C}]} \tag{17}$$

Considering (17) the closed loop transfer function of the system according to deviation from the balancing loop has the form

$$\Phi_{k\varepsilon}(s) = \Phi_\varepsilon(s)(1 - e^{-sT_c}) \tag{18}$$

Figure 4 shows Bode characteristics of a standard closed-loop system according to disorder with according to disorder with the compensation loop $Lk(\omega)$ and without compensation loop $L(\omega)$

Complex s form of the transfer function of the synchronous filter does not allow the use of traditional methods of stability analysis. Therefore, to obtain the stability conditions we analyze the transfer function of the compensation component

$$H_k(s) = \frac{\mathrm{H}(s) \cdot e^{-sT_c}}{(1 - \beta \cdot e^{-sT_c})} \tag{19}$$

Where the transfer function of the position control system is normalized to the frequency of intersection ω=1

$$H(s) = \frac{\gamma \cdot (1 + 3{,}5s) \cdot e^{-sT_c}}{(1 - \beta \cdot e^{-sT_c}) \cdot (1 + 3{,}5s + 3{,}3s^2 + 2s^3)} \tag{20}$$

Where in the system is stable if hodograph $H(s)$ is within a circle of radius $1/\gamma$. Fig.5. shows influence the parameters β and γ to the area of stability in the plane

$H(s)$. The graphs of the compliance compensation process were calculated for different γ coefficients and parametersβ, which are implemented in the form of filter.

$$\beta(s) = 1/(1+s) \tag{21}$$

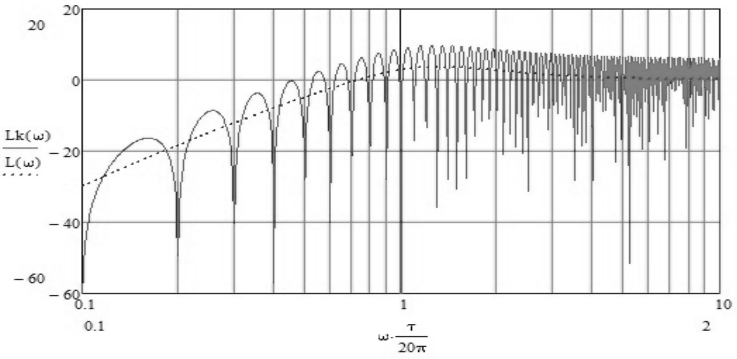


Fig. 4. Bode diagram of a standard closed-loop system under disturbance

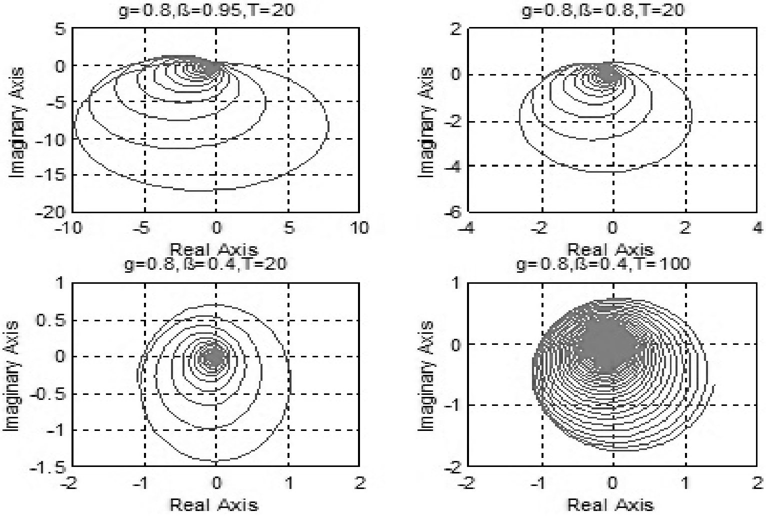


Fig. 5. Graphs of the process of convergence of compensation

4 Conclusion

The paper deals with the issue of development of the systems to stabilize the angular velocity with high precision. Maximum accuracy of the stabilization of the stabilization is limited by the work zone control system. Delimitation of the work zone is a complex technical problem, the solution of which is affected by a large number of factors. In consideration of substantive issues, the authors propose the following sequence of solutions of synthesis control system: For a given object structure, the required transfer function of the control loop is selected. For the required transfer function is recommended the use of the 3-1-2-4 characteristic type, which is optimized by setting criteria of symmetric optimum». In the second phase, the problem of coupling of desired characteristics to the real object parameter control is solved. Depending on the construction of the drive there are proposed certain methods of selecting the intersection frequency of the desired characteristic. Analysis of steady-state values also shows that the convergence process of compensation is determined by placing hodograph in the transmission function of compensation within a circle with a radius of one.

References

1. Isermann, R.: Digital Control Systems. Fundamentals, Deterministic Control, vol. 1. Springer, Berlin (1989)
2. Gabriska, D., Michalconok, G., Tanuska, P., Skulavik, T.: Analisys of the compensation algorithm satability of disturbance for the phase control system. In: 11th IFAC/IEEE International Conference on Progammable Devices and Embedded Systems (PDeS 2012), May 23-24, pp. 100–104. BRNO (2012)
3. Gabriska, D.: Analiz schodimosti procesa kompensacie osibki regulirovania. In: Aktuaľnyje Problemy i Innovacii v Ekonomike, Upravlenii, Obrazovanii, Informacionnych Technologijach: Materialy Meždunarodnoj Naučnoj Konferencii, Stavropoľ-Kislovodsk, Maja 3-7, vol. 6(1), pp. 177–180. NOU VPO CevKavGTI, Stavropoľ (2011) ISSN 2074-1685
4. Chemodanov, B.I.: Sledjaschie provoda. Kniga pervaja, M., Energia (1976)
5. O'dwaer, A.: Handbook of PI and PID Controller Tuning Rules. Imperial College Press, Singapore (2003) ISBN: 13 978-1-84816-242-6
6. Levine, W.S.: The control handbook. CRC Press (2000)

Polling System with Threshold Control for Modeling of SIP Server under Overload

Sergey Shorgin[1], Konstantin Samouylov[2], Yuliya Gaidamaka[2], and Shamil Etezov[2]

[1] Institute of Informatics Problems of RAS, Moscow, Russia
sshorgin@ipiran.ru
[2] Peoples' Friendship University of Russia, Moscow, Russia
{ksam,ygaidamaka}@sci.pfu.edu.ru, setezov@gmail.com

Abstract. The main purpose of this research is the development of the methods for realizing an overload control mechanism on the Session Initiation Protocol servers by the application of polling systems with different service disciplines. The mathematical model is studied by means of numerical methods of the queuing theory and allows analyzing the behavior of different control parameters depending on the load in network of SIP servers. The polling system consists of two queues of finite capacity and implements the threshold control of loading by low-priority customers. The exhaustive and gated service disciplines are studied under Markov assumptions; formulas for calculation of the main probability measures of the polling system are derived. By performing simulations we demonstrate that the polling system with a threshold in the priority queue is a possible solution for loss-based overload control scheme at the SIP server. In some cases from the viewpoint of server utilization we found that the gated discipline is more preferable.

Keywords: polling system, exhaustive discipline, gated discipline, overload control, SIP server, threshold control.

1 Introduction

Session Initiation Protocol (SIP) is the prevailing signaling protocol for full range of services such as Voice-over-IP, presence service, instant messaging, video conferencing, and multimedia distribution [1]. Wide expansion of SIP surfaces the problem of overload in SIP-server networks that can occur for many reasons, e.g. capacity planning, component failures, flash crowds, denial of service attacks [2]. The built-in SIP overload control mechanism based on generating rejection response messages (503 Service Unavailable) does not solve the problem and can even worsen the situation by propagating the overload and bringing potential network collapse. Some solutions have been proposed to prevent SIP overload for different degree of cooperation (hop-by-hop, end-to-end or local overload control), explicit and implicit overload control schemes [3, 4]. A number of papers are devoted to the development of SIP server models, both analytical and simulation

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_10, © Springer International Publishing Switzerland 2014

[5–19]. We investigate one of the explicit overload control feedback mechanism namely Loss-Based Overload Control (LBOC) [4]. According to LBOC scheme when a SIP server experiences overload it notifies its upstream neighboring SIP server about overloading and specifies the percent of the messages going downstream that should be discarded or redirected by the upstream SIP server.

In this paper we study the polling system [20] with two queues of finite capacity and threshold control of low-priority customers' load, as a model of SIP server overload control. In the Markov assumptions exhaustive and gated service disciplines are studied, the method for calculation of infinitesimal matrix of the Markov process is proposed, formulas for calculation of the main probability measures are derived.

The contribution of this paper is two-fold. First, the model takes into account the message prioritization recommended by the IETF [3]. In session establishment the *Invite* request has low priority versus any *nonInvite* response since the last conduces successive session establishment. Second, the model considers the threshold control with a threshold in the high priority queue and two polling service disciplines, exhaustive and gated. In some cases the latter is preferable.

The paper is organized as follows. The second section contains detailed description of the model for both service disciplines. Markov process and its state space are introduced. The transitions intensity diagram for exhaustive service discipline, which is simpler in representation, is presented. In the third section an equilibrium equations are derived, and the statements, allowing to form the block three-diagonal matrix for applying of known numerical methods, e.g. LU decomposition, and finding a steady-state probabilities distribution, are made. The fourth section contains the numerical analysis of the main probability measures, i.e. average number of customers in the system, probability of system being in the normal loading states as well as in the overloading states, and the sever utilization. Finally the tasks for further research are formulated.

2 Model Description

Let's consider a mathematical model of the SIP server functioning as a system with high-priority type 1 customers, low-priority type 2 customers and a single server carrying out cyclic poll of queues with finite capacity $r_1 < \infty$ and $r_2 < \infty$, $r_1 + r_2 = R$, where R is a server's buffer capacity. Flows of customers arriving on the system are Poisson with intensity values λ_1 and λ_2 respectively and duration of service has the exponential distribution with parameters μ_1 and μ_2 respectively. Customers in each queue are served in FCFS order. The threshold $0 < L < r_1$ is introduced in the 1-st queue for overload control. When the threshold is reached the system goes to the overload mode and flow intensity of low-priority type 2 customers decreases to $\lambda_2' < \lambda_2$ value. Switching time between two queues is distributed under the exponential law with parameter s_1 when switching from the second queue to the first and with parameter s_2 when switching from the first queue to the second one. It is assumed that the serving customer holds its place in the queue until leaving the system.

Exhaustive Service Discipline

According to exhaustive service discipline the server processes customers of the queue until the number of customers drops to zero and then it switches to the next queue. Let $l(t) \in \{1, 2\}$ be the number of the queue serving by the server at the moment $t > 0$, and $n_l(t)$ is the number of customers in l-th queue at the moment $t > 0$. Let's define the stochastic process $\eta(t) = (l(t),\ n_1(t),\ n_2(t))$ describing functioning of the polling system with exhaustive discipline which is the Markov process with the state-space

$$\mathcal{X} = \{(l, n_1, n_2)|\ l \in \{1, 2\},\ n_1 \in \{0, ..., r_1\},\ n_2 \in \{0, ..., r_2\}\}, \qquad (1)$$

and $|\mathcal{X}| = 2\,(r_1 + 1)\,(r_2 + 1)$.

The state-space X breaks into two disjoint subspaces:

$$\mathcal{X}_0 = \{(l, n_1, n_2)|\ l \in \{1, 2\},\ n_1 \in \{0, ..., L\},\ n_2 \in \{0, ..., r_2\}\} \qquad (2)$$

and

$$\mathcal{X}_1 = \{(l, n_1, n_2)|\ l \in \{1, 2\},\ n_1 \in \{L + 1, ..., r_1\},\ n_2 \in \{0, ..., r_2\}\}. \qquad (3)$$

In Fig. 1 the transitions diagram is depicted as well as its space splitting into two subspaces of normal and overload states is shown.

Gated Service Discipline

According to gated service discipline the server processes only customers which were in the queue at the moment of server's connection to the queue, and the customers which arrived in the queue after server's connection will be served in the following cycle.

Let's denote $m(t)$ a number of type l customers which is needed to be served by the server in the current cycle at the moment $t > 0$. Then the stochastic process $\eta(t) = (l(t), n_1(t), n_2(t), m(t))$ describing functioning of the system with gated discipline is a Markov process with the state-space

$$\mathcal{X} = \left\{ \begin{array}{l} (l, n_1, n_2, m):\ m \in \{0, ..., \max(r_1, r_2)\};\ n_l \in \{m, ..., r_l\}, \\ n_{l \bmod 2+1} \in \{0, ..., r_{l \bmod 2+1}\};\ l = 1, 2 \end{array} \right\}, \qquad (4)$$

and $|\mathcal{X}| = (r_1 + 1)\,(r_2 + 1)\left(\frac{r_1+r_2+4}{2}\right)$.

Hereinafter, if it will not require additional explanations, we will use the same notations for system with both service disciplines. Then, as well as for exhaustive discipline, we will break the state-space $\mathcal{X}$ of the Markov process $\{\eta(t),\ t \geq 0\}$ into two disjoint subspaces:

$$\mathcal{X}_0 = \left\{ \begin{array}{l} (l, n_1, n_2, m) : l \in \{1; 2\}; m \in \{0, ..., r_l\}, n_l = \{m, ..., r_l\}, \\ n_{l \bmod 2\ +1} \in \{0, ..., r_{l \bmod 2\ +1}\}, n_1 \leq L \end{array} \right\} \qquad (5)$$

$$\mathcal{X}_1 = \left\{ \begin{array}{l} (l, n_1, n_2, m) : l \in \{1; 2\}; m \in \{0, ..., r_l\}, n_l = \{m, ..., r_l\}, \\ n_{l \bmod 2\ +1} \in \{0, ..., r_{l \bmod 2\ +1}\}, n_1 > L \end{array} \right\}. \qquad (6)$$

In the next section the method of Markov process states ordering for both disciplines is offered so that its infinitesimal matrix could be presented in a block three-diagonal form for numerical solution of the equilibrium equations through the UL decomposition method.

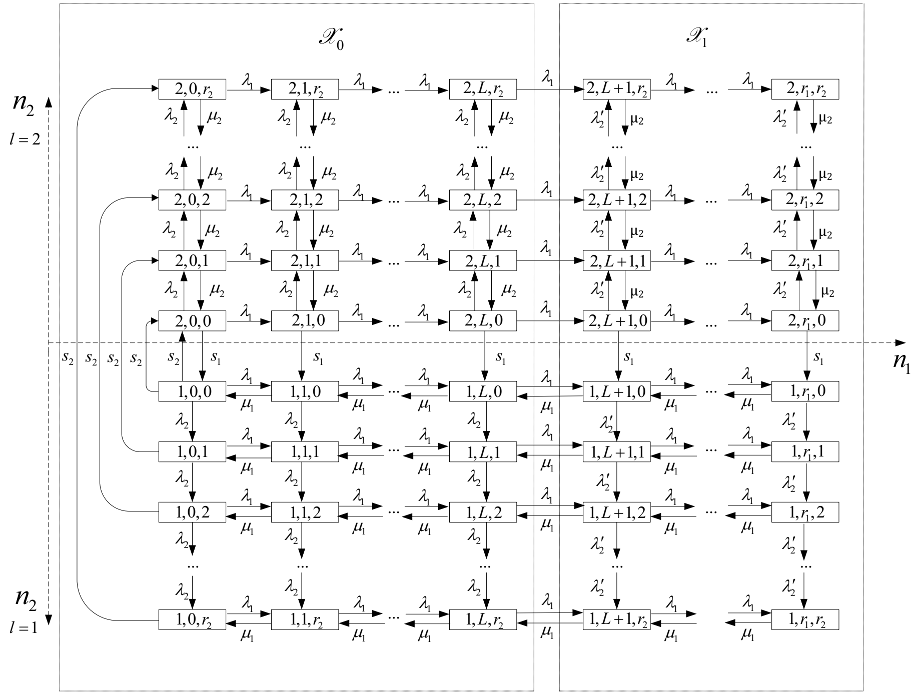


Fig. 1. Transition diagram of the Markov process $\{\eta(t),\ t \geq 0\}$ for exhaustive discipline

3 Equilibrium Equations

Let's denote $p(l, n_1, n_2)$ a steady-state probability of the system with exhaustive discipline being in the state $(l, n_1, n_2) \in \mathcal{X}$ and then the equilibrium equations take the following form:

$$\begin{aligned}
&(s_2 u\,(1-n_1) + \lambda_1 u\,(r_1-n_1) + \mu_1 u(n_1) + \lambda_2 u\,(r_2-n_2)\,u\,(L-n_1+1) + \\
&+\lambda_2' u\,(r_2-n_2)\,u\,(n_1-L))p\,(1,n_1,n_2) = \\
&= s_1 u\,(1-n_2)\,p\,(2,n_1,n_2) + \mu_1 u\,(r_1-n_1)\,p\,(1,n_1+1,n_2) + \\
&+\lambda_2 u\,(n_2)\,u\,(L-n_1+1)\,p\,(1,n_1,n_2-1) + \\
&+\lambda_2' u\,(n_2)\,u\,(n_1-L)\,p\,(1,n_1,n_2-1) + \lambda_1 u\,(n_1)\,p\,(1,n_1-1,n_2)\,, \\
&n_1 \in \{0,...,r_1\}\,,\ n_2 \in \{0,...,r_2\}\,,\ l=1, \\
&\\
&(s_1 u\,(1-n_2) + \lambda_1 u\,(r_1-n_1) + \mu_2 u\,(n_2) + \\
&+\lambda_2 u\,(r_2-n_2)\,u\,(L-n_1+1) + \lambda_2' u\,(r_2-n_2)\,u\,(n_1-L))p\,(2,n_1,n_2) = \\
&= s_2 u\,(1-n_1)\,p\,(1,n_1,n_2) + \mu_2 u\,(r_2-n_2)\,p\,(2,n_1,n_2+1) + \\
&+\lambda_1 u\,(n_1)\,p\,(2,n_1-1,n_2) + \lambda_2 u\,(n_2)\,u\,(L-n_1+1)\,p\,(2,n_1,n_2-1) + \\
&+\lambda_2' u\,(n_2)\,u(n_1-L)p\,(2,n_1,n_2-1)\,, \\
&n_1 \in \{0,...,r_1\}\,,\ n_2 \in \{0,...,r_2\}\,,\ l=2,
\end{aligned} \tag{7}$$

where $u(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases}$

We now denote $p(l, n_1, n_2, m_l)$ a steady-state probability of the system with gated discipline being in state $(l, n_1, n_2, m) \in \mathcal{X}$ and representing its equilibrium equations as following:

$$\begin{array}{l}
(s_2 u(1-m) + \lambda_1 u(r_1 - n_1) + \lambda_2 u(r_2 - n_2) u(L + 1 - n_1))\, p_{1,n_1,n_2,m} + \\
+ (\lambda_2' u(r_2 - n_2) u(n_1 - L) + \mu_1 u(m))\, p_{1,n_1,n_2,m} = \\
= \lambda_1 u(n_1 - m) p_{1,n_1-1,n_2,m} + \lambda_2 u(n_2) u(L + 1 - n_1) p_{1,n_1,n_2-1,m} + \\
+ \lambda_2' u(n_2) u(n_1 - L) p_{1,n_1,n_2-1,m} + \mu_1 u(r_1 - n_1) p_{1,n_1+1,n_2,m+1} + \\
+ s_1 u(m + 1 - n_1) p_{2,n_1,n_2,0}, \\
n_2 \in \{0, \, ... \, , r_2\}, \;\; m \in \{0, \, ... \, , r_1\}, \;\; n_1 \in \{m, \, ... \, , r_1\}; \\
\\
(s_1 u(1-m) + \lambda_1 u(r_1 - n_1) + \lambda_2 u(r_2 - n_2) u(L + 1 - n_1))\, p_{2,n_1,n_2,m} + \\
+ (\lambda_2' u(r_2 - n_2) u(n_1 - L) + \mu_2 u(m))\, p_{2,n_1,n_2,m} = \\
= \lambda_1 u(n_1) p_{2,n_1-1,n_2,m} + \lambda_2 u(n_2) u(L + 1 - n_1) p_{2,n_1,n_2-1,m} + \\
+ \lambda_2' u(n_2 - m) u(n_1 - L) p_{2,n_1,n_2-1,m} + \mu_2 u(r_2 - n_2) p_{2,n_1,n_2+1,m+1} + \\
+ s_2 u(m + 1 - n_1) p_{1,n_1,n_2,0}, \\
n_1 \in \{0, \, ... \, , r_1\}, \; m \in \{0, \, ... \, , r_2\}, \; n_2 \in \{m, \, ... \, , r_2\}.
\end{array} \tag{8}$$

The solution of equilibrium equations (7) and (8) can be obtained numerically and to do this requires to calculate their matrix $\mathbf{A}$ – the infinitesimal matrix of Markov process $\{\eta(t), \; t \geq 0\}$. For this purpose we introduce the following lexicographic order on the state-space $\mathcal{X}$ and we consider that:

for exhaustive discipline $\mathbf{x} = (l, n_1, n_2) \prec \mathbf{x}' = (l', n_1', n_2')$, if

$$\begin{array}{l}
(n = n_1 + n_2 < n' = n_1' + n_2') \vee \\
\vee \{(n = n') \wedge [(l < l') \vee ((l = l') \wedge (n_l < n_{l'}'))]\},
\end{array} \tag{9}$$

for gated discipline $\mathbf{x} = (l, n_1, n_2, m) \prec \mathbf{x}' = (l', n_1', n_2', m')$, if

$$\begin{array}{l}
(n < n') \vee \\
\vee \left\{ (n = n') \wedge \left[(l < l') \vee \left((l = l') \wedge \left((m < m') \vee \begin{pmatrix} (m = m') \wedge \\ \wedge (n_l < n_{l'}') \end{pmatrix} \right) \right) \right] \right\}.
\end{array} \tag{10}$$

Statement 1. If the lexicographic order (9) for the exhaustive discipline and (10) for the gated discipline are introduced on the state-space $\mathcal{X}$ then:

1. Infinitesimal matrix $\mathbf{A}$ for both service disciplines can be represented in a block three-diagonal form

$$\mathbf{A} = \begin{pmatrix}
D_0 & UP_0 & 0 & & 0 & 0 \\
LW_1 & D_1 & UP_1 & \cdots & 0 & 0 \\
0 & LW_2 & D_2 & & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & \cdots & D_{R-1} & UP_{R-1} \\
0 & 0 & 0 & & LW_R & D_R
\end{pmatrix}. \tag{11}$$

2. The non-zero off-diagonal elements of blocks D, UP, LW of matrix **A** for the exhaustive discipline can be calculated by the following formulas:

$$D_n = (d_{M(l,n_1,n_2),M(l',n_1,n_2)}),$$
$$l,l' = 1,2,\ n_1+n_2 = n,\ n = 0,...,R,$$
$$D_n = \begin{cases} s_1,\ n_2 = 0,\ l = 2,\ l' = 1, \\ s_2,\ n_1 = 0,\ l = 1,\ l' = 2, \\ 0,\ \text{in other cases,} \end{cases}$$
$$UP_n = (u_{M(l,n_1,n_2),M(l,n'_1,n'_2)}),$$
$$l = 1,2,\ n_1+n_2 = n'_1+n'_2-1 = n,\ n = 0,...,R-1,$$
$$UP_n = \begin{cases} \lambda_1,\ n_1 = n'_1-1,\ n_2 = n'_2, \\ \lambda_2,\ n_1 = n'_1,\ n_2 = n'_2-1,\ n_1 < L+1, \\ \lambda'_2,\ n_1 = n'_1,\ n_2 = n'_2-1,\ n_1 \geq L+1, \\ 0,\ \text{in other cases,} \end{cases}$$
$$LW_n = (lw_{M(l,n_1,n_2),M(l,n'_1,n'_2)}),$$
$$l = 1,2,\ n_1+n_2 = n'_1+n'_2+1 = n,\ n = 1,...,R,$$
$$LW_n = \begin{cases} \mu_1,\ l = 1,\ n_1 = n'_1+1,\ n_2 = n'_2, \\ \mu_2,\ l = 2,\ n_1 = n'_1,\ n_2 = n'_2+1, \\ 0,\ \text{in other cases,} \end{cases}$$

where the element (l,n_1,n_2) of the array $M(l,n_1,n_2)$ contains index of the state (l,n_1,n_2) in the matrix **A**.

3. The non-zero off-diagonal elements of blocks D, UP, LW of matrix **A** for gated discipline can be calculated by the following formulas:

$$D_n = (d_{M(l,n_1,n_2,m),M(l',n_1,n_2,m')}),$$
$$l,l' = 1,2,\ n_1+n_2 = n,\ n = 0,...,R,$$
$$D_n = \begin{cases} s_1,\ n_2 = 0,\ m = 0,\ l = 2,\ l' = 1,\ m' = n_1, \\ s_2,\ n_1 = 0,\ m = 0,\ l = 1,\ l' = 2,\ m' = n_2, \\ 0,\ \text{in other cases,} \end{cases}$$
$$UP_n = (u_{M(l,n_1,n_2,m),M(l,n'_1,n'_2,m)}),$$
$$l = 1,2,\ n_1+n_2 = n'_1+n'_2-1 = n,\ n = 0,...,R-1,$$
$$UP_n = \begin{cases} \lambda_1,\ n_1 = n'_1-1,\ n_2 = n'_2, \\ \lambda_2,\ n_1 = n'_1,\ n_2 = n'_2-1,\ n_1 < L+1, \\ \lambda'_2,\ n_1 = n'_1,\ n_2 = n'_2-1,\ n_1 \geq L+1, \\ 0,\ \text{in other cases,} \end{cases}$$
$$LW_n = (lw_{M(l,n_1,n_2,m),M(l,n'_1,n'_2,m')}),$$
$$l = 1,2,\ n_1+n_2 = n'_1+n'_2+1 = n,\ n = 1,...,R,$$
$$LW_n = \begin{cases} \mu_1,\ l = 1,\ n_1 = n'_1+1,\ n_2 = n'_2,\ m = m'+1, \\ \mu_2,\ l = 2,\ n_1 = n'_1,\ n_2 = n'_2+1,\ m = m'+1, \\ 0,\ \text{in other cases,} \end{cases}$$

where the element (l,n_1,n_2,m) of the array $M(l,n_1,n_2,m)$ contains index of the state (l,n_1,n_2,m) in the matrix **A**.

4 Numerical Analysis

Matrix **A** representation in the form of (11) allows solving equilibrium equations numerically, e.g. using the UL decomposition method. Knowing probability distribution $p\left(l, n_1, n_2\right)$ for the system with exhaustive discipline we can calculate an average number of customers in queues using the formulas

$$N_1 = \sum_{n_1=1}^{r_1} \left(n_1 \cdot \sum_{n_2=1}^{r_2} \sum_{l=1}^{2} p\left(l, n_1, n_2\right) \right)$$

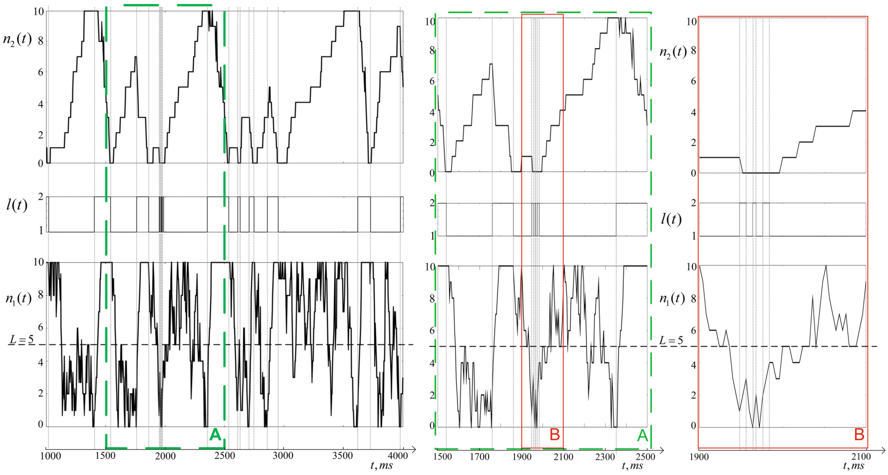


Fig. 2. The stochastic processes $\{l(t), n_1(t), n_2(t), t > 0\}$ for exhaustive discipline

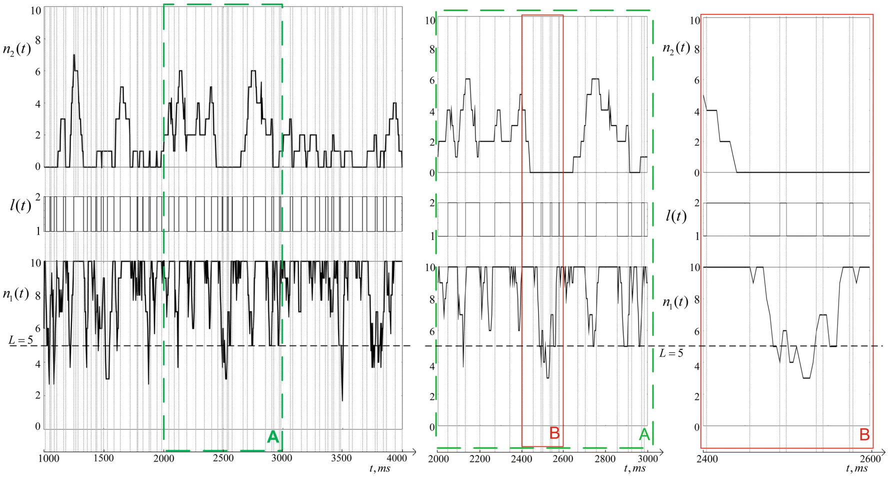


Fig. 3. The stochastic processes $\{l(t), n_1(t), n_2(t), t > 0\}$ for gated discipline

and

$$N_2 = \sum_{n_2=1}^{r_2} \left(n_2 \cdot \sum_{n_1=1}^{r_1} \sum_{l=1}^{2} p\left(l, n_1, n_2\right) \right).$$

The probabilities of the system being in the normal load states and $\mathcal{X}_0$ in the overload states $\mathcal{X}_1$ can be calculated using the formulas

$$P(\mathcal{X}_0) = \sum_{(l,n_1,n_2)\in\mathcal{X}_0} p\left(l, n_1, n_2\right) \quad \text{and} \quad P(\mathcal{X}_1) = \sum_{(l,n_1,n_2)\in\mathcal{X}_1} p\left(l, n_1, n_2\right)$$

respectively. The utilization factor of the server can be found as $U = 1 - P_0$, where

$$P_0 = \sum_{n_2=0}^{r_2} p(1, 0, n_2) + \sum_{n_1=0}^{r_1} p(2, n_1, 0).$$

Probability measures of the system with gated discipline are calculated in a similar way.

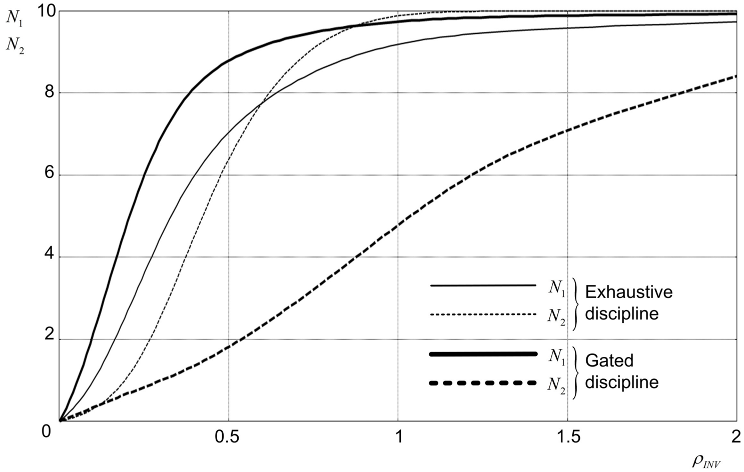


Fig. 4. Average number of customers in the system

While carrying out the numerical analysis we suppose that typical session of the user requires transfer of one *Invite* message and six *nonInvite* messages [21], i.e. $\lambda_1 = 6\lambda_2$. For simplicity of calculations we assume $R = 20$, $r_1 = 10$, $r_2 = 10$, $L = 5$, $\lambda_2' = 0,5\lambda_2$. It is supposed that processing time of the *Invite* messages is $\mu_2^{-1} = 10$ msec and the *nonInvite* messages is $\mu_1^{-1} = 4$ msec. The switching time between queues is comparable to these values and is equal to $s_1^{-1} = s_2^{-1} = 10$ msec.

In Fig. 2, Fig. 3 and Fig. 4 the comparison of the probability measures is presented. Graphs are depicted depending on the load created by *Invite* messages

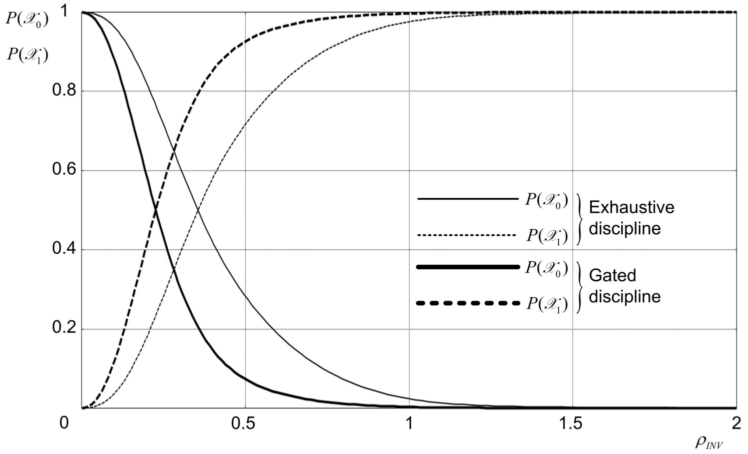


Fig. 5. Probability of finding the system in normal loading states and in overload states

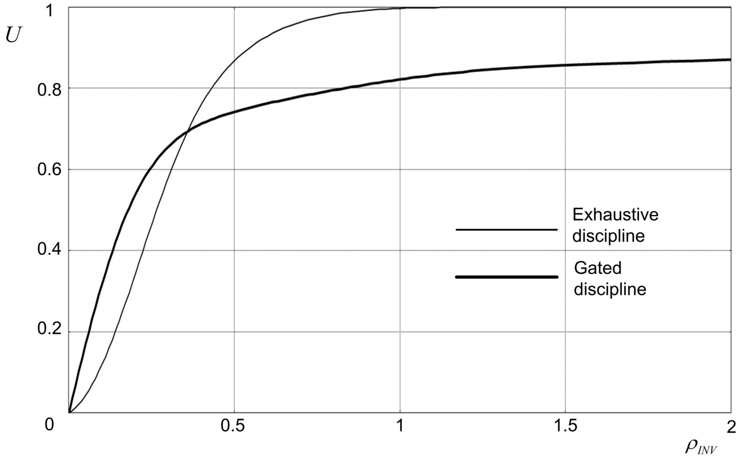


Fig. 6. The utilization factor of the server

i.e. $\rho_{INV} = \rho_2 = \lambda_2 \cdot \mu_2^{-1}$ value. Since we study the overload case the chosen range of loading values is $0 \leq \rho_{INV} \leq 2$. In our case from the point of view of server utilization we received the gated discipline preferable.

Fig. 5 and 6 illustrate the stochastic processes $\{l(t), n_1(t), n_2(t), t > 0\}$ behavior on the time interval [1000msec; 4000msec] for exhaustive and gated disciplines respectively. We assume that the low-priority queue loading is $\rho_{INV} = 0.4$. These graphs are calculated using the matrix exponential representation

$$P(t) = \sum_{n=0}^{\infty} \frac{(t \cdot \mathrm{A})^n}{n!} = e^{\mathrm{A}t}$$

of the transition probability matrix of the stochastic process $\{\eta(t), t > 0\}$.

5 Conclusion

SIP server overload control is of a great importance nowadays. First reason is that the cost of rejecting a signaling message is not negligible compared to the cost of serving a user request for session establishment. Second, SIP has a client-server architecture which helps the development of threshold control solution at different priority schemes.

We proposed the polling system with two queues and with a threshold in the priority queue as a solution for the LBOC scheme described in [4]. We developed the simulation model of the polling system and obtained the numerical method for performance evaluation of main control parameters of SIP-server. Finally, some case study was made in order to compare two possible service disciplines – the exhaustive discipline and the gated discipline. We also illustrated the polling system behavior in transient mode in order to show the features of its functioning.

Our future work may include development of the simulation framework for analysis of SIP server behavior under overloading. The most interesting problem is minimization the return time from overloading to the normal operation.

Acknowledgements. This work was supported in part by the Russian Foundation for Basic Research (grant 12-07-00108-a and 13-07-00665-a).

References

1. Rosenberg, J., Schulzrinne, H., Camarillo, G., et al.: SIP: Session Initiation Protocol. IETF RFC 3261 (2002)
2. Rosenberg, J.: Requirements for Management of Overload in the Session Initiation Protocol. IETF RFC 5390 (2008)
3. Hilt, V., Noel, E., Shen, C., Abdelal, A.: Design Considerations for Session Initiation Protocol (SIP) Overload Control. IETF RFC 6357 (2011)
4. Gurbani, V., Hilt, V., Schulzrinne, H.: Session Initiation Protocol (SIP) Overload Control (2013); draft-ietf-soc-overload-control-12
5. Ohta, M.: Overload Protection in a SIP Signaling Network. In: International Conference on Internet Surveillance and Protection, pp. 205–210 (2006)
6. Noel, E.C., Johnson, C.R.: Initial simulation results that analyze SIP based voIP networks under overload. In: Mason, L.G., Drwiega, T., Yan, J. (eds.) ITC 2007. LNCS, vol. 4516, pp. 54–64. Springer, Heidelberg (2007)
7. Hilt, V., Widjaja, I.: Controlling Overload in Networks of SIP Servers. In: IEEE International Conference on Network Protocols, pp. 83–93 (2008)
8. Shen, C., Schulzrinne, H., Nahum, E.: Session Initiation Protocol (SIP) Server Overload Control: Design and Evaluation. In: Schulzrinne, H., State, R., Niccolini, S. (eds.) IPTComm 2008. LNCS, vol. 5310, pp. 149–173. Springer, Heidelberg (2008)
9. Montagna, S., Pignolo, M.: Performance Evaluation of Load Control Techniques in SIP Signaling Servers. In: Proceedings of Third International Conference on Systems (ICONS), pp. 51–56 (2008)
10. Noel, E.C., Johnson, C.R.: Novel Overload Controls for SIP Networks. In: Proceedings of 21st International Teletraffic Congress (2009)

11. Ohta, M.: Overload Control in a SIP Signaling Network. International Journal of Electrical and Electronics Engineering, 87–92 (2009)
12. Garroppo, R.G., Giordano, S., Spagna, S., Niccolini, S.: Queueing Strategies for Local Overload Control in SIP Server. In: IEEE Global Telecommunications Conference, pp. 1–6 (2009)
13. Sun, J., Tian, R.X., Hu, J.F., Yang, B.: Rate-based SIP Flow Management for SLA Satisfaction. In: Proceedings of 11th International Symposium on Integrated Network Management (IEEE/IFIP IM), New York, USA, pp. 125–128 (2009)
14. Homayouni, M., Nemati, H., Azhari, V., Akbari, A.: Controlling Overload in SIP Proxies: An Adaptive Window Based Approach Using No Explicit Feedback. In: IEEE Global Telecommunications Conference GLOBECOM 2010, pp. 1–5 (2010)
15. Hong, Y., Huang, C., Yan, J.: Analysis of SIP Retransmission Probability Using a Markov-Modulated Poisson Process Model. In: Proceedings of IEEE/IFIP Network Operations and Management Symposium, Osaka, Japan, pp. 179–186 (2010)
16. Abdelal, A., Matragi, W.: Signal-Based Overload Control for SIP Servers. In: Proceedings of IEEE CCNC, Las Vegas, NV (2010)
17. Abaev, P., Gaidamaka, Y., Samouylov, K.E.: Queuing Model for Loss-Based Overload Control in a SIP Server Using a Hysteretic Technique. In: Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART 2012. LNCS, vol. 7469, pp. 371–378. Springer, Heidelberg (2012)
18. Abaev, P., Gaidamaka, Y., Pechinkin, V., Razumchik, R., Shorgin, S.: Simulation of overload control in SIP server networks. In: Proceedings of the 26th European Conference on Modelling and Simulation, ECMS 2012, Germany, Koblenz, pp. 533–539 (2012)
19. Abaev, P., Gaidamaka, Y., Samouylov, K., Pechinkin, V., Razumchik, R., Shorgin, S.: Hysteretic control technique for overload problem solution in network of SIP servers. To appear in Computing and Informatics 1 (2014)
20. Vishnevsky, V., Semenova, O.: Polling Systems: Theory and Applications for Broadband Wireless Networks, 317 p. Lambert Academic Publishing, London (2012)
21. Johnston, A., Donovan, S., Sparks, R., et al.: Session Initiation Protocol (SIP) Basic Call Flow Examples. IETF RFC 3665 (2003)

The Control Moment Gyroscope Inverted Pendulum

Yawo H. Amengonu[1], Yogendra P. Kakad[1], and Douglas R. Isenberg[2]

[1] Electrical and Computer Engineering Department,
University of North Carolina at Charlotte,
Charlotte, NC 28223
{yhameng1,kakad}@uncc.edu
[2] Aerospace and Mechanical Engineering Department,
Embry-Riddle Aeronautic University
Prescott, AZ 86301
isenberd@erau.edu

Abstract. This paper presents the dynamics and control of a control moment gyroscope actuated inverted pendulum. The control technique utilizes partial feedback linearization, an appropriate global change of coordinates which transform the dynamics of the system into a lower nonlinear subsystem and a chain of double integrators. A control Lyapunov function is designed in order to stabilize the overall system using a backstepping procedure.

Keywords: Underactuated nonlinear systems, Lagrangian Dynamics, Systems with kinematic symmetry, Backstepping procedure, Partial feedback linearization.

1 Introduction

There are perhaps only a few simple systems that are better than the inverted pendulum at demonstrating the ability to accomplish a seemingly difficult task through the use of feedback control. It is therefore no surprise that the inverted pendulum has been extensively utilized as a prototype system for both the study and practical demonstration of many types of controllers. Inverted pendulum systems are often attached to a cart or rotating arm that in which case the angle of the pendulum is controlled via the coupling between the translational motion of the pendulum's pivot point and its angle, a consequence of the conservation of momentum. An interesting variation on this problem is the momentum wheel inverted pendulum [1],[2],[3]. In this case, the pendulum's pivot point is inertially fixed and actuation is accomplished via the controlled rotation of a massive disk attached to the pendulum about an axis parallel to the pendulum's pivot axis. In this paper, a similar configuration is utilized, however, the massive rotating disk is allowed to also rotate about an axis that is parallel to the length of the pendulum. Such a mechanism forms a simple control moment gyroscope (CMG) which can be utilized to provide a torque on the pendulum. The dynamics of

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_11, © Springer International Publishing Switzerland 2014

the control moment gyroscope are first presented. The equations of motion are derived from the Euler-Lagrange formulation. From these equations, some statements can be made concerning the control requirements. These are addressed. A stabilizing controller for the pendulum is then examined.

2 System Dynamics

Consider the diagram of the control moment gyroscope inverted pendulum that is depicted in figure 1. There are three bodies in this system, body 1, body 2, and body 3. There are also three degrees-of-freedom associated with the pendulum's rotation through angle θ_1, the CMG's rotation through angle θ_2, and the CMG disk's rotation through angle θ_3. A vector of generalized coordinates for this system is thus, $\underline{\gamma} = [\theta_1, \theta_2, \theta_3]^T$, the subscript T stands for transpose.The system dynamics are easily obtained from Hamilton's principle via the Euler-Lagrange equation. However, this requires an expression for the total system kinetic energy and potential energy. The total kinetic energy is obtained as the sum of the kinetic energies of each of the bodies comprising the system,

$$K = \sum_{B=1}^{3} K_B \ , \tag{1}$$

where

$$K_B = \frac{1}{2} m_B \underline{{}^I\dot{R}_B^T} \underline{{}^I\dot{R}_B} + \underline{{}^I\dot{R}_B^T} {}^I\dot{T}_B \underline{\Gamma_B} + \frac{1}{2} \underline{{}^B\omega_B}^T J_B \underline{{}^B\omega_B} \ . \tag{2}$$

The potential energy of the system is obtained as

$$U = \sum_{B=1}^{3} U_B \ , \tag{3}$$

where

$$U_B = g \left(m_B \underline{{}^I R_B} + {}^I T_B \underline{\Gamma_B} \right) |_z \ . \tag{4}$$

In both of these expressions, m_B is the mass of the body, $\underline{\Gamma_B}$ is the vector of first-mass moments of the body measured in the body's frame, J_B is the inertial matrix of the body measured in the body's frame, ${}^I T_B$ is the rotation from body coordinates to inertial coordinates, $\underline{{}^B\omega_B}$ is the angular velocity of body measured in the body's frame, $\underline{{}^I R_B}$ is the position of the body's frame measured in the inertial frame, and g is the gravitational acceleration.

Formulating the Lagrangian as $L = K - U$ and applying the Euler-Lagrange equation results in a system of second-order differential equations of the form

$$H\left(\underline{\gamma}\right) \underline{\ddot{\gamma}} + \underline{D}\left(\underline{\gamma}, \underline{\dot{\gamma}}\right) + \underline{G}\left(\underline{\gamma}\right) = \underline{\tau} \ , \tag{5}$$

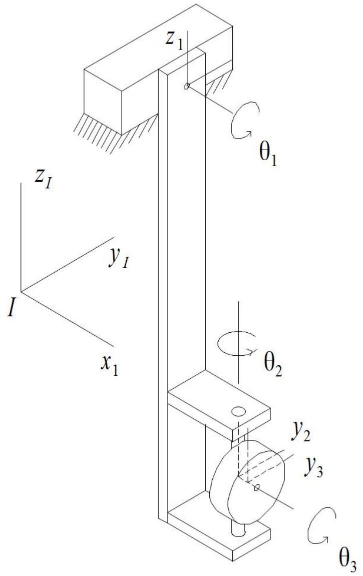


Fig. 1. Control Moment Inverted Pendulum

where the components of the symmetric positive-definite system mass matrix, $H\left(\underline{\gamma}\right)$, are

$$\begin{aligned} H_{1,1} &= a + b\sin(\theta_2)^2 \\ H_{1,2} &= H_{2,1} = c\cos(\theta_2) \\ H_{1,3} &= H_{3,1} = J_{3xx}\cos(\theta_2) \\ H_{2,2} &= d \\ H_{2,3} &= H_{3,2} = 0 \\ H_{3,3} &= J_{3xx} \quad . \end{aligned} \tag{6}$$

The components of the vector of the generalized Coriolis and centripetal forces, $\underline{D}\left(\underline{\gamma}, \underline{\dot{\gamma}}\right)$, are

$$\begin{aligned} D_1 &= 2b\dot{\theta}_1\dot{\theta}_2\cos(\theta_2)\sin(\theta_2) + J_{3xx}\dot{\theta}_2\dot{\theta}_3\sin(\theta_2) - c\dot{\theta}_2^{\,2}\sin(\theta_2) \\ D_2 &= -b\dot{\theta}_1^{\,2}\cos(\theta_2)\sin(\theta_2) + J_{3xx}\dot{\theta}_1\dot{\theta}_3\sin(\theta_2) \\ D_3 &= -J_{3xx}\dot{\theta}_1\dot{\theta}_2\sin(\theta_2) \quad . \end{aligned} \tag{7}$$

The components of the vector of generalized gravitational forces, $\underline{G}\left(\underline{\gamma}\right)$, are

$$\begin{aligned} G_1 &= Ag\sin(\theta_1) + Bg\cos(\theta_1)\sin(\theta_2) \\ G_2 &= Bg\sin(\theta_1)\cos(\theta_2) \\ G_3 &= 0 \quad , \end{aligned} \tag{8}$$

where

$$\begin{aligned}
a &= J_{1xx} + J_{2xx} + J_{3xx} + {d_1}^2 m_3 \\
b &= J_{3yy} - J_{3xx} + d_3^2 m_3 \\
c &= d_1 d_3 m_3 \\
d &= m_3 d_3^2 + J_{2zz} + J_{3yy} \\
A &= d_1 m_2 - \Gamma_{1z} + d_1 m_3 \\
B &= d_3 m_3 \ .
\end{aligned} \tag{9}$$

In (6)-(8), d_1 is the distance along the z-axis of frame 1 from the origin of frame 1 to the origin of frame 2, d_2 is the distance along the x-axis of frame 1 from the origin of frame 1 to the origin of frame 2, and d_3 is the distance from the origin of frame 2 to the origin of frame 3.

3 Dynamics Analysis

The input $\underline{\tau}$ is a vector of external generalized forces, $\underline{\tau} = [0, \tau_2, \tau_3]^T$. The system is therefore under-actuated. The torque τ_3 is utilized to maintain a constant $\dot{\theta}_3$, thus only τ_2 is available for the control of θ_2.

Assuming a control has been applied to regulate $\dot{\theta}_3$ about a setpoint such that $\ddot{\theta}_3 = 0$, the equations of motion are approximated as

$$\begin{aligned}
H_{1,1}(\theta_2)\ddot{\theta}_1 + H_{1,2}(\theta_2)\ddot{\theta}_2 + D_1\left(\underline{\theta}, \underline{\dot{\theta}}\right) + G_1\left(\underline{\theta}\right) &= 0 \\
H_{2,1}(\theta_2)\ddot{\theta}_1 + H_{2,2}(\theta_2)\ddot{\theta}_2 + D_2\left(\underline{\theta}, \underline{\dot{\theta}}\right) + G_2\left(\underline{\theta}\right) &= \tau_2 \ .
\end{aligned} \tag{10}$$

3.1 Collocated Partial Feedback Linearization

Many researchers, in the past, have considered the analysis and control design of underactuated mechanical systems. One of the complexities of these systems is that often they are not fully feedback linearizable. In this work we will partially linearize the system using a change of control which transforms it into a strict feedback form[4] and then into a normal form which is a special case of the famous Byrnes-Isidori normal form [5]. This form is suitable the backstepping procedure. However, after applying this change of control, the new control appears in both the linear and nonlinear subsystems. Another change of variable renders the analysis and design less complicated because the control appears only in the linear subsystem. The global change of control was proposed in [6] as

$$\begin{aligned}
\tau_2 &= \alpha\left(\underline{\theta}\right) u + \beta\left(\underline{\theta}, \underline{\dot{\theta}}\right) \\
\ddot{\theta}_2 &= u \ .
\end{aligned} \tag{11}$$

Equation (11) is obtained by solving for $\ddot{\theta}_1$ in the first equation in (10) and then replace it in the second equation. Applying this technique, (10) is partially linearized and we have

$$\begin{aligned}\alpha(\theta_2) &= \left(H_{2,2}(\theta_2) - \frac{H_{1,2}(\theta_2)H_{2,1}(\theta_2)}{H_{1,1}(\theta_2)}\right)\\ \beta(\underline{\theta},\underline{\dot{\theta}}) &= D_2(\underline{\theta}) + G_2(\underline{\theta}) - \frac{H_{1,2}(\theta_2)}{H_{1,1}(\theta_2)}\left(D_1(\underline{\theta}) + G_1(\underline{\theta})\right) \ .\end{aligned}$$

The reduced system (10) may be written as

$$\begin{aligned}\ddot{\theta}_1 &= -\frac{1}{H_{1,1}(\theta_2)}(D_1(\underline{\theta},\underline{\dot{\theta}}) + G_1(\underline{\theta})) - \frac{1}{H_{1,1}(\theta_2)}H_{1,2}(\theta)u\\ \ddot{\theta}_2 &= u \ . \end{aligned} \tag{12}$$

In general this change of control is invertible because $det\,(H(\underline{\theta}))$ is not zero because $H(\underline{\theta})$ is a positive definite matrix. In (10), $H_{1,1}(\theta_2)$ is positive for all values of θ_2. Denoting $\underline{p} = \underline{\dot{\theta}}$, (12) can be expressed as

$$\begin{aligned}\dot{\theta}_1 &= p_1\\ \dot{p}_1 &= f(\underline{\theta},\underline{p}) + g(\underline{\theta})u\\ \dot{\theta}_2 &= p_2\\ \dot{p}_2 &= u \ .\end{aligned} \tag{13}$$

In (13), it is shown that the control input u appears in both the (θ_1,p_1) and (θ_2,p_2) subsystems. It is interesting to note that the inertia matrix depends only on the actuated variable. The variables that appear in the inertia matrix are called shape variables. The configuration variables that do not appear in the inertia matrix are called external variables. The notion of shape or internal variables originally appeared in the control literature by the study of multi-body systems [7],[8].

3.2 Shape Variable and Kinetic Symmetry

The fact that the Lagrangian has a kinematic symmetry with respect to external variable i.e $\frac{\partial K(\theta,\dot{\theta})}{\partial \theta_j} = 0$, $j = 1, 3$, and the normalized momentum conjugate to θ_1,

$$\nu_1 = H_{1,1}(\theta_2)^{-1}\frac{\partial L}{\partial \dot{\theta}_1} = \dot{\theta}_1 + H_{1,1}(\theta_2)^{-1}H_{1,2}(\theta_2)\dot{\theta}_2 \tag{14}$$

is integrable. This defines an interesting group action in the configuration manifold. This action is defined in a more general way in [9].
Let

$$\psi(\theta_2) = \int_0^{\theta_2} \frac{H_{1,2}(s)}{H_{1,1}(s)}\,ds \ ,$$

we note that the one-form $d\psi(\theta_2) = \frac{H_{1,2}(\theta_2)}{H_{1,1}(\theta_2)}d(\theta_2)$ is exact and the above fact is exploited to perform a global change of coordinates as

$$\begin{aligned}z_1 &= \theta_1 + \psi(\theta_2)\\ z_2 &= H_{1,1}(\theta_2)\dot{\theta}_1 + H_{1,2}(\theta_2)\dot{\theta}_2 = \frac{\partial L}{\partial \dot{\theta}_1} \ .\end{aligned} \tag{15}$$

The above global change of variables transform the dynamics of the reduced system into a strict feedback form as

$$\begin{aligned}
\dot{z_1} &= \frac{z_2}{H_{1,1}(\theta_2)} \\
\dot{z_2} &= g_1(z_1 - \psi(\theta_2), \theta_2) \\
\dot{\theta}_2 &= p_2 \\
\dot{p_2} &= u \ ,
\end{aligned} \tag{16}$$

where

$$g_1(z_1 - \psi(\theta_2), \theta_2) = \frac{d}{dt}\frac{\partial L}{\partial \dot{\theta_1}} = \frac{\partial L}{\partial \theta_1} = -\frac{\partial U}{\partial \theta_1} \ ,$$

due to kinetic symmetry with respect to θ_1, $\frac{\partial K}{\partial \theta_1} = 0$. We also note that this change of variables is possible because $H_{1,1}(\theta_2)$ is strictly positive for all values of θ_2 and by multiplying (14) by $H_{1,1}(\theta_2)$ and setting $y_1 = z_1$ and $y_2 = {H_{1,1}}^{-1}(\theta_2) z_2$, one obtains a special case of the normal form [5] with double integrators as shown in (17)

$$\begin{aligned}
\dot{y_1} &= y_2 \\
\dot{y_2} &= f(y, \zeta_1, \zeta_2) \\
\dot{\zeta_1} &= \zeta_2 \\
\dot{\zeta_2} &= u \ .
\end{aligned} \tag{17}$$

Clearly, the control input appears only in the actuated subsystem. This decouples the two subsystems with respect to the control input u. If a globally stabilizing smooth state feedback exists for (z_1, z_2)-subsystem in (16) then a globally stabilizing state feedback can be found for (θ_1, θ_2)-subsystem using backstepping procedure [4]. In this case, θ_2 is considered as virtual input connecting both subsystems.

4 Controller Design Using Backstepping

Inertia-wheel pendulum is a planar inverted pendulum with a revolving wheel at the end that was first introduced in [10]. Due to the fact all the Christoffel Symbols associated with the inertia matrix vanish and the inertia matrix is constant, the dynamics and control of the inertia-wheel pendulum is a particular case of the design procedure outlined in this paper.

Let us first consider the stabilization of the nonlinear (z_1, z_2)-subsystem

$$\begin{aligned}
\dot{z_1} &= H_{1,1}(\theta_2)^{-1} z_2 \\
\dot{z_2} &= g_1(z_1 - \psi(\theta_2), \theta_2) \ ,
\end{aligned} \tag{18}$$

where

$$\psi(\theta_2) = \frac{c}{\sqrt{ab}} \arctan\left(\frac{b\sin(\theta_2)}{\sqrt{ab}}\right) \ , \tag{19}$$

with $\theta_2 \in (-\frac{\pi}{2}\frac{\pi}{2})$. Clearly, the subsystem (18) is non-affine in the virtual control input θ_2. The stabilization of (18) can be done using the following assumption.

Consider the above nonlinear system non-affine in control θ_2 in (18). If the following condition

$$g_1(z_1, \theta_2) = \frac{\partial L}{\partial \theta_1}$$

is a smooth function with $g_1(0,0) = 0$, $H_{1,1}(\theta_2) > 0$ for all values of θ_2, zero is not a critical value for $g_1(z_1, \theta_2)$ and $\frac{\partial g_1(z_1,\theta_2)}{\partial \theta_2} \neq 0$, on the manifold $M = ker(g_1) = \{(z_1, \theta_2) \in \mathbb{R}^2 : g_1(z_1, \theta_2) = 0\}$ and $g_1(z_1, \theta_2)$ has an isolated root $\alpha(z_1)$ such that $g_1(z_1, \alpha(z_1)) = 0$, so there exists a continuously differentiable state feedback law in the following form $\theta_2 = \alpha(z_1) - \sigma(z_1, z_2)$ that globally asymptotically stabilizes the origin of (18) ($\sigma(.)$ is a sigmoidal function. Refer to [9] for proof).

$$\begin{aligned} g_1(z_1, \theta_2) &= Ag\sin(z_1 - \psi(\theta_2)) + Bg\cos(z_1 - \psi(\theta_2))\sin(\theta_2) = 0 \\ \frac{\partial g_1(z_1, \theta_2)}{\partial \theta_2} &= -Ag\frac{H_{1,2}(\theta_2)}{H_{1,1}(\theta_2)}\cos(z_1 - \psi(\theta_2)) + Bg\frac{H_{1,2}(\theta_2)}{H_{1,1}(\theta_2)}\sin(z_1 - \psi(\theta_2)\sin(\theta_2) \\ &+ Bg\cos(z_1 - \psi(\theta_2))\cos(\theta_2) \ . \end{aligned} \tag{20}$$

Solving the first line in (20) is equivalent to first solving for θ_1 in $Ag\sin(\theta_1) + Bg\cos(\theta_1)\sin(\theta_2) = 0$ which gives

$$\theta_1 = -\arctan\left(\frac{B\sin(\theta_2)}{A}\right) \ , \tag{21}$$

and substituting it in the first line of (15) we have

$$z_1 = \frac{c}{\sqrt{ab}} \arctan\left(\frac{b\sin(\theta_2)}{\sqrt{ab}}\right) - \arctan\left(\frac{B\sin(\theta_2)}{A}\right) \ , \tag{22}$$

with $\theta_2 \in (-\frac{\pi}{2}\frac{\pi}{2})$. In our work, we numerically invert (22) and find θ_2 as a function of z_1 . From here a state feedback control law stated in the assumption above can be found to stabilize the (z_1, z_2)-subsystem to the origin (0,0) as

$$\theta_2 = K_1(z_1, z_2) = \alpha(z_1) - \sigma(c_1 z_1 + c_2 z_2) \ . \tag{23}$$

4.1 Backstepping Control Design

Let us define a control Lyapunov fuction $V(z_1, z_2, \eta_1)$ and $\eta_1 = \theta_2 - K_1(z_1, z_2)$ as

$$\begin{aligned} V(\eta_1) &= \frac{1}{2}{\eta_1}^2 \\ \dot{V}(\eta_1) &= \dot{\eta_1}\eta_1 \ , \end{aligned} \tag{24}$$

where

$$\begin{aligned} \dot{V}(\eta_1) &= \dot{\eta_1}\eta_1 \\ \dot{\eta_1} &= -c_3\eta_1 \\ p_2 - \dot{K}_1(z_1, z_2) &= -c_3(\theta_2 - K_1(z_1, z_2)) \\ &\text{or} \\ p_2 &= K_2(z_1, z_2, \theta_2), \end{aligned}$$

or

$$p_2 = K_2(z_1, z_2, \theta_2) = -c_3(\theta_2 - K_1(z_1, z_2)) + \dot{K}_1(z_1, z_2),$$

where $c_3 > 0$. Next we define

$$\begin{aligned} \eta_2 &= p_2 - K_2(z_1, z_2, \theta_2) \\ V(z_1, z_2, \eta_1, \eta_2) &= \frac{1}{2}{\eta_1}^2 + \frac{1}{2}{\eta_2}^2 \\ \dot{V}(z_1, z_2, \eta_1, \eta_2) &= -c_3{\eta_1}^2 + \dot{\eta_2}\eta_2 \\ \dot{\eta_2} &= -c_4\eta_2 \\ u - \dot{K}_2(z_1, z_2, \theta_2) &= -c_4\left(p_2 - K_2(z_1, z_2, \theta_2)\right), \end{aligned}$$

or

$$u = K_3(z_1, z_2, \theta_2, p_2) = -c_4\left(p_2 - K_2(z_1, z_2, \theta_2)\right) + \dot{K}_2(z_1, z_2, \theta_2) \ , \tag{25}$$

where $c_4 > 0$.

With $\dot{\theta}_3$ regulated about a setpoint such that $\ddot{\theta}_3 = 0$, the overall system can be controlled using τ_2 which is available for the control of θ_2.

5 Simulation

The simulation result for the overall dynamics of the Control Moment Gyroscope is shown in figure 2. An oscillation was noticed around the equilibrium point. To overcome this issue, a nonlinear damping was added. The values used for simulation are shown in the table 1.

Table 1. Simulation parameters

a	b	c	d	J_{311}	A	B	c_1	c_2	c_3	c_4
2694.6	6.7728	75.5573	7.3916	6.8707	879.6338	22.9022	1	0	10	8

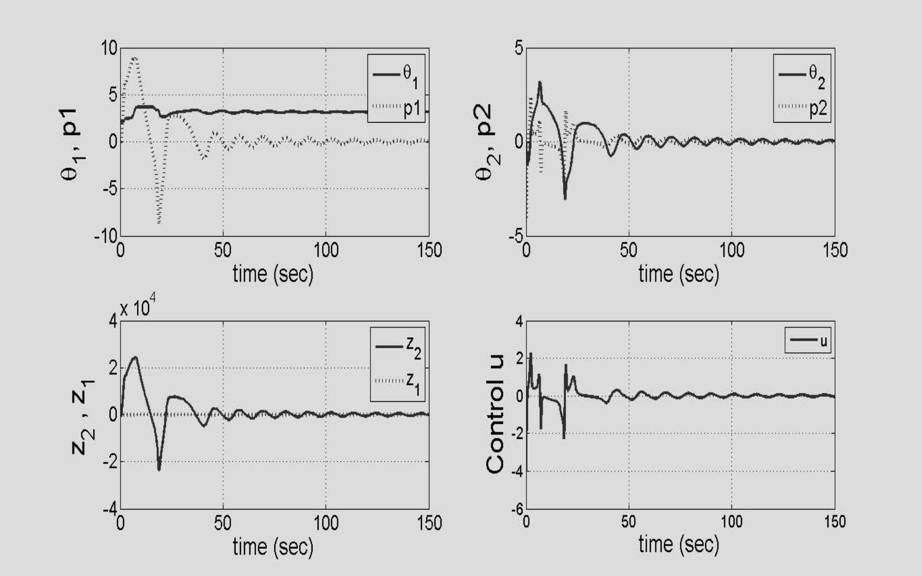


Fig. 2. Simulation result with initial conditions $[-\frac{\pi}{3}, 0, 0, 0]$, where $p_1 = \dot{\theta}_1, p_2 = \dot{\theta}_2$

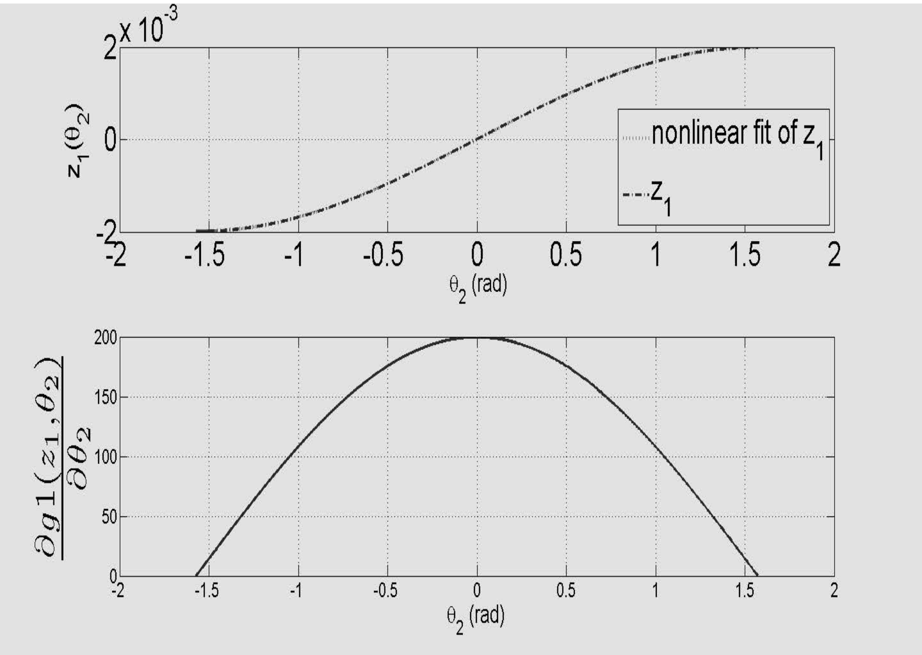


Fig. 3. Analysis of $z_1(\theta_2)$ and $\frac{\partial g1(z_1,\theta_2)}{\partial\theta_2}$

6 Conclusion and Future Work

We consider the dynamics analysis and control of the control moment gyroscope inverted pendulum. The CMG is an underactuated system and has kinematic symmetry with respect of some of its space configuration variables. The system is first partial feedback linearized. Then, we use the kinematic symmetry of the system to perform a global change of coordinates which transform the original system into a lower order nonlinear subsystem plus a chain of double integrators. The backstepping procedure is used to stabilize the cascade systems and the original system is stabilized at one of its unstable equilibrium points.It was noticed that the system becomes unstable when θ_2 falls outside $(-\frac{\pi}{2}, \frac{\pi}{2})$. We propose in future to find a change of coordinates such that the system is stablizable for $\theta_2 \in [-\pi, \pi]$.

References

[1] Spong, M.W.: Partial feedback linearization of underactuated mechanical systems. In: Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems, IROS 1994, Advanced Robotic Systems and the Real World, vol. 1. IEEE (1994)

[2] Spong, M.W.: The swing up control problem for the acrobot. IEEE Control Systems Magazine (1995)

[3] Spong, M.W.: Underactuated mechanical systems. In: Siciliano, B., Valavanis, K.P. (eds.) Control Problems in Robotics and Automation. LNCIS, vol. 230, pp. 135–150. Springer, Heidelberg (1998)

[4] Krstić, M. Kanellakopoulos, I., Kokotović, P.: Nonlinear and Adaptive Control Design. John Wiley & Sons (1995)

[5] Isidoris, A.: Nonlinear Control systems. Springer (1995)

[6] Spong, M.W.: Energy Based Control of a Class of Underactuated Mechanical Systems. In: 1996 IFAC World Congress (July 1996)

[7] Krishnaprasad, P.S., Simo, J.: Dynamics and Control of Multi-body Systems, vol. 97. AMS, Providence (1989)

[8] Krishnaprasad, P.S.: Geometric phases and optimal reconfiguration for multi-body systems. In: Proc. of American Control Conference, San Diego, CA (1990)

[9] Olfati-Saber, R.: Nonlinear Control of Underactuated Mechanical Systems with Application to Robotics and Aerospace Vehicles. PhD thesis, Massachusetts Institute of Technology, Department Electrical and Computer, Science (February 2001)

[10] Spong, M.W., Corke, P., Lozano, R.: Nonlinear control of the Inertia Wheel Pendulum. Automatica (September 1999)

Realization of an Inverted Pendulum Robot Using Nonlinear Control for Heading and Steering Velocities

Danielle S. Nasrallah, Sylvain Brisebois, and Maarouf Saad

Department of Electrical Engineering,
École de technologie supérieure,
Montreal, Canada
{danielle.nasrallah,sylvain.brisebois}@gmail.com,
maarouf.saad@etsmtl.ca

Abstract. This paper describes the realization of InPeRo, an *In*verted *Pe*ndulum *Ro*bot pertaining to the class of *mobile wheeled pendulums* (MWP), and the implementation of a nonlinear controller responsible of controlling the heading velocity of the robot and its steering rate while stabilizing its central body.

The mathematical model of the robot is formulated first followed by the equations governing the controller. Then the construction of the robot is described including the selection of components and instruments on board. The implementation of the controller equations, data acquisition and analysis, filtering, calibration, sensor fusion and other issues are addressed. Finally, experimental results are shown to validate the effectiveness of the controller.

Keywords: Mobile wheeled pendulum, Realization, Instrumentation, Nonlinear control implementation, Experimental validation.

1 Introduction

This paper describes the realization of a wheeled inverted pendulum robot InPeRo (edition 2) and the implementation of a nonlinear controller responsible of controlling the heading velocity and the steering rate while stabilizing the central body. InPeRo pertains to the class of mobile wheeled pendulums. MWP-class robots are composed of three rigid bodies: two wheels rotating about a central body. A feature common to MWPs, that is not encountered in other wheeled robots, is that their central body, which constitutes the robot platform, can rotate about the wheel axis. This motion must be controlled, thereby leading to a new challenging problem for MWP, which is the *stabilization of the central body*, aside the classical control problem due to nonholonomy. Additionally, depending on the position of the centre of mass of the central body with respect to the line of wheel centres (above/below) the MWP robot can be classified as inverted or non-inverted.

J. Świątek et al. (eds.), *Advances in Systems Science*, Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_12, © Springer International Publishing Switzerland 2014

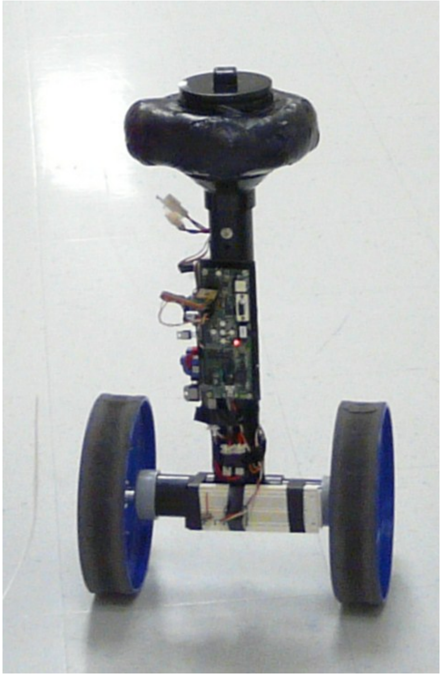


Fig. 1. InPeRo, an inverted pendulum robot, Courtesy of É.T.S

Many developments in the field of MWP have been reported since 1999: The US patent behind the Ginger and then the Segway Human Transporter projects [1]; JOE, a mobile inverted pendulum [2]; QuasiMoRo, a quasiholonomic mobile robot [3]; ATOM, an anti-tilting outdoor mobile robot [4,5]; uBot, a dynamically balancing two-wheeled platform [6]; nBot, a two-wheeled balancing robot [7]; The first edition of InPeRo [8]; Winglet, a personal transport assistant robot from Toyota [9]; Recon Scout Throwbot, the reconnaissance robot for military applications from Recon Robotics [10]; I-PENTAR, an inverted pendulum type assistant robot [11]; and, more recently, a mobile inverted pendulum robot system [12].

Segway, Winglet and ReconScout being three commercial products with high budgets invested on development we will not compare our work to them. However, it is noteworthy to highlight the emergence of MWPs and their penetration into the market. We will rather concentrate on laboratory products with similar development budgets. As mentioned previously the stabilization of the central body is a must in MWPs, hence, the measure and control of the tilt angle and its time rate-of-change are essential. In JOE three sensors have been evaluated for this purpose: (i) a tilt-sensor, (ii) an accelerometer and (iii) a gyro. The latter was retained, however it involved integration drift. Additionally, the controller was based on decoupling and linearization. The whole performance of the system was acceptable but the authors suggested an adaptive controller for better results. In QuasiMoRo a solid-state sensor was used for tilt measurement combined to a linear controller, however, experimental results were not given. As for ATOM, the work was limited to the simulation level and no prototype was developed. In uBot a combination of gyro and accelerometer was employed together with a LQR regulator. The design appeared to be successful, nevertheless the authors suggested many adjustments for later improvements. In the first edition of InPeRo two optical distance measuring sensors were mounted on a 90−degrees fixture to measure the tilt angle in a combination with a nonlinear controller. The results were satisfactory but the authors found many aspects that need to be improved, they will be listed subsequently. In the work done by Lee and Jung a fusion of a gyro and a tilt-sensor was employed. The scheme involved three filters: (i) a high-pass filter for the gyro, (ii) a low-pass filter for the tilt, and

(iii) a Kalman filter for the fusion. Linear PD and PID regulators were used for heading angle and position control. The results were successful, however the authors suggested advanced control algorithms to cope with tracking errors and achieve better performance.

In the work presented here we use a simple scheme for the fusion of the tilt and gyro sensors together with a nonlinear controller. Aside the successful tracking, the results reported show robustness versus parameters uncertainties, external disturbances and change of the position of the centre of mass of the central body. The novelty of the work is summarized below:

1. A simple scheme for the fusion of two distance-measuring sensors and a gyro resulting in a fast reliable reading.
2. A nonlinear controller composed of three imbricated loops responsible of current (torque) control of DC-brush motors, stabilization of the central body, and heading velocity and steering rate control.
3. A decentralization of the control scheme and the data acquisition rendering the whole design modular and easy to debug and repair.
4. A robustness versus parameters uncertainties, external disturbances, live-change of the position of the centre of mass of the central body.
5. A *light hardware* and *do-it-yourself* design approach that requires low budget and high creativity.

On the paper organization, Section II describes the mathematical model of the robot and the equations governing the controller. The construction of the robot including the material used are given in Section III. The choice of actuators and sensors, setup and calibration are described in Section IV. Finally Section V shows the experimental results and validates the performance of the controller.

2 Mathematical Model and Nonlinear Control

The mathematical model of the robot has been developed in a previous work and considered the motion on an inclined plane. We will give here the final results, the reader being directed to [4] for the details.

MWP being composed of three rigid bodies, the wheels are denoted bodies 1 and 2, while the central body is body 3. The centre of mass of the wheels coincide with the geometric centre while the centre of mass of the central boy is located above the line of wheel centres, rendering InPeRo an inverted pendulum robot. The symbols used for the robot modelling are summarized in Table 1.

2.1 State-Space Formulation

The posture of the robot can be described by a six-dimensional vector $\mathbf{q}$ defined as

$$\mathbf{q} = \begin{bmatrix} x_{\mathbf{c}_o} \; y_{\mathbf{c}_o} \; \mathbf{r}^T \; r_0 \end{bmatrix}^T \tag{1}$$

The three-dimensional vector of independent generalized velocities is

$$\mathbf{v} = \begin{bmatrix} v_c \; \omega_{3p} \; \omega_{3l} \end{bmatrix}^T \tag{2}$$

Table 1. List of Symbols

b	Distance between the wheel centres
$\mathbf{c}_i$	Position vector of the center of mass C_i of the i^{th} body, $i = 1, 2, 3$
$\mathbf{c}_o$	Position vector of C_o the midpoint of wheels centres
d	Offset between C_o and C_3
$\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$	Right-handed orthogonal triad describing the orientation of the inertial frame $\mathcal{F}_0$
$\mathbf{l}$	Unit vector along the line of wheel centers, directed from C_1 to C_2
m_c	Mass of the central body
m_w	Mass of each wheel
r_w	Radius of each wheel
r_b	Radius of the base cylinder
r_0, $\mathbf{r}$	The Euler-Rodrigues parameters describing the orientation of the central body in $\mathcal{F}_0$
$\{\mathbf{u}_i, \mathbf{l}, \mathbf{v}_i\}$	Right-handed orthogonal triad describing the orientation of the i^{th} body
$\mathbf{v}_3$	Unit vector directed from C_3 to C_o
$\mathbf{I}_c$	Inertia matrix of the central body
$\mathbf{I}_w$	Inertia matrix of each wheel
θ_{i3}	Angular displacement of the i^{th} wheel w.r.t. the central body
τ_i	Input torque applied to the i^{th} wheel
$\boldsymbol{\omega}_i$	Angular velocity vector of the i^{th} body in $\mathcal{F}_0$

where v_c is the heading velocity of the robot, namely,

$$v_c = \frac{r_w}{2}(\dot{\theta}_{13} + \dot{\theta}_{23} + 2\omega_{3l})$$

while ω_{3p} and ω_{3l} represent the projection of the angular velocity vector of the central body $\boldsymbol{\omega}_3$ along the vertical $\mathbf{k}$ and the line of wheel centres $\mathbf{l}$, respectively,

$$\boldsymbol{\omega}_3 = \omega_{3p}\mathbf{k} + \omega_{3l}\mathbf{l}$$

It is noteworthy that ω_{3p} is the steering rate of the robot, namely,

$$\omega_{3p} = \frac{r_w}{b}(\dot{\theta}_{13} - \dot{\theta}_{23})$$

The nine-dimensional state vector thus becomes

$$\mathbf{x} = \begin{bmatrix} \mathbf{q}^T & \mathbf{v}^T \end{bmatrix}^T \tag{3}$$

The computation of $\dot{\mathbf{x}}$ requires the kinematic constraints and the robot dynamics. The former are given by

$$\dot{\mathbf{c}}_o = v_c\mathbf{h},\ \dot{\mathbf{r}} = \frac{1}{2}\left(r_0\mathbf{1} - \mathbf{R}\right)\boldsymbol{\omega}_3, \text{ and } \dot{r}_0 = -\frac{1}{2}\mathbf{r}^T\boldsymbol{\omega}_3$$

where $\mathbf{h}$ is a unit vector given by $\mathbf{h} = \mathbf{l} \times \mathbf{k}$ and $\mathbf{R}$ is the cross-product matrix (CPM)[1] of vector $\mathbf{r}$, and $\mathbf{1}$ is the 3×3 identity matrix.

[1] The CPM of a vector $\mathbf{v} \in \mathbb{R}^3$ is defined, for every $\mathbf{x} \in \mathbb{R}^3$ as well, as $\mathbf{V} = \text{CPM}(\mathbf{v}) = \partial(\mathbf{v} \times \mathbf{x})/\partial\mathbf{x}$.

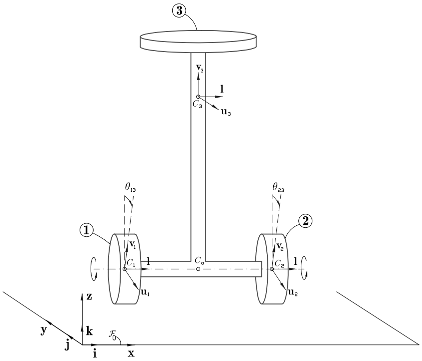


Fig. 2. InPeRo, an inverted pendulum robot

The dynamics, developed using the Natural Orthogonal Complement method, yields

$$\dot{\mathbf{v}} = \mathbf{I}^{-1}(\mathbf{q})\left[-\mathbf{C}(\mathbf{q},\mathbf{v})\mathbf{v} + \boldsymbol{\tau}(\mathbf{q}) + \boldsymbol{\gamma}(\mathbf{q})\right]$$

where $\mathbf{I}$ and $\mathbf{C}$ are 3×3 matrices representing the generalized inertia of the system and the contribution of the Coriolis and centrifugal forces, respectively, while $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ are three-dimensional vectors representing the generalized active and gravity forces, respectively.

Consequently, the full state-space model of the system becomes:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}_p(r_0, \mathbf{r})\tau_p + \mathbf{g}_m(r_0, \mathbf{r})\tau_m \tag{4}$$

with

$$\tau_p = \tau_1 + \tau_2 \text{ and } \tau_m = \tau_1 - \tau_2$$

For more details on the expressions of $\mathbf{f}(\mathbf{x})$, $\mathbf{g}_p$ and $\mathbf{g}_m$, see [4,5].

2.2 Nonlinear Controller

The nonlinear controllability study and a nonlinear controller have been developed in a previous work. We will give here the final results, the reader being directed to [5] for the details. The control scheme is given in Fig. 3 below:

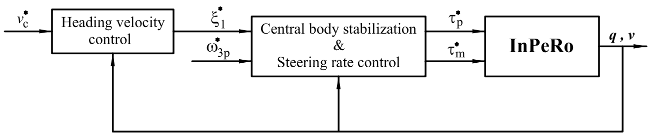


Fig. 3. Nonlinear control scheme

3 Robot Skeleton Material

As mentioned previously a *light hardware* and *do-it-yourself* design approach have been employed in the construction of InPeRo. The skeleton of the robot used plumbing parts essentially straight and T-shaped pipes. The wheels are composed of two frisbees collated together as shown in Fig. 4.

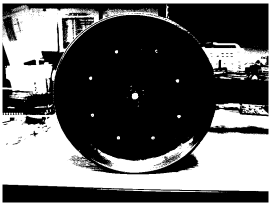


Fig. 4. Frisbee Wheel

4 Actuators and Sensors

4.1 Motors, Gearboxes and Encoders

The input of a MWP-class robot are the torques applied to the wheels. In the case of InPeRo two DC-brush motors are used. The reason behind the selection of DC-brush motors is that the relationship current/torque is linear, hence, the motor torque control problem becomes a simple current control problem. The motors are selected from the RE-max series of Maxon, namely 226774. Motors specifications are summarized in Table 2.

Additionally, the gearboxes and encoders are also from Maxon, 144035 and 225771 respectively, and have the following specifications:

– Gearbox: planetary gearhead, 53:1
– Encoder: quadratic encoder, 3 channels, 128 counts

Table 2. Motor Specifications

Power	9 Watt
Nominal Voltage	24 Volt
Nominal Current	0.568 Amp
Nominal Speed	4080 rpm
Nominal Torque	25.9 mNm
Speed Constant	206 rpm/Volt
Torque Constant	46.3 mNm/Amp

4.2 Torque Control

As mentioned above, controlling the torque of a DC-brush motor is a matter of controlling its current. In the first edition of InPeRo a linear PID controller was implemented in the microcontroller. The performance was acceptable first, however, a degradation was noticed due to online changes of the electrical parameters (resistance and inductance) and an adjustment of the PID parameters was required constantly.

For that, in the current edition of InPeRo we decided to use a hysteresis current control (HCC) that is robust with respect to parameters uncertainties. Additionally, for modularity purposes and in order to enlighten the code inside the microcontroller the HCC is implemented on a separate card that communicates with the microcontroller via a DAC using SPI communication. The scheme showing the HCC principle is given in Fig. 5.

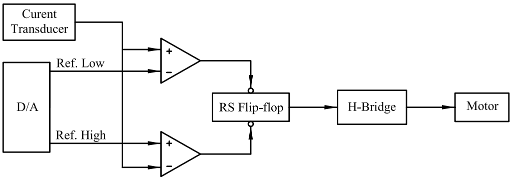


Fig. 5. Hysteresis Current Control

The components used for the HCC card are:

- Current Tranducer: LEM LTS 6-NP
- D/A converter: TLV 5604
- Fast dual operational amplifier: LM2903
- Dual D-type flip-flop: SN74AHC74
- H-Bridge: LMD18200

4.3 Tilt-Angle Measurement

The tilt angle and its time rate-of-change are measured using the fusion of two distance measuring sensors with a gyroscope. Two GP2D120, for distance measuring, are mounted on a 90-degrees fixture on the base cylinder as shown in Fig. 6.

Let δ_b and δ_f denote the backward and forward distances of the sensors to the ground, respectively. By the same token, let α_b and α_f represent the angles of the backward and forward sensors with the ground, respectively. The tilt angle α is thus obtained as $alpha = 45^o - \alpha_b$ with

$$\alpha_b = \text{atan}\left(\frac{r_b + l_s + \delta_f}{r_b + l_s + \delta_b}\right)$$

where l_s is the sensor length.

The angle reading was tested, it is accurate, however, the delay between two readings can go up to 48ms (38.3 ± 9.6ms precisely). For that we found the necessity of introducing another type of angle measuring sensor to increase reading fastness and robustness. The gyroscope IDG300 was mounted along the line of wheel centres to measure the time rate-of-change of the tilt-angle. However, as it is well-known, the numerical integration to obtain the angle introduces drift. For that a continuous correction at the output of the integrator was adopted. The scheme showing the fusion of the two distance measuring sensors and the gyroscope is given in Fig. 7.

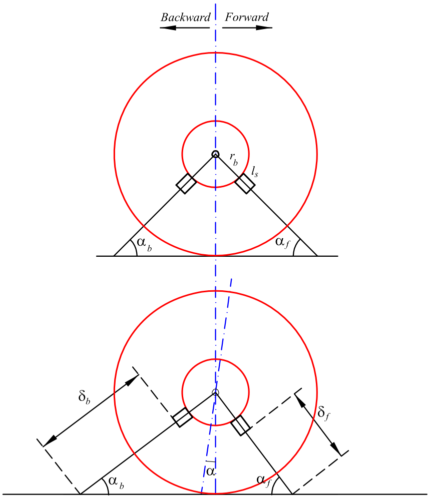


Fig. 6. Distance measuring sensors

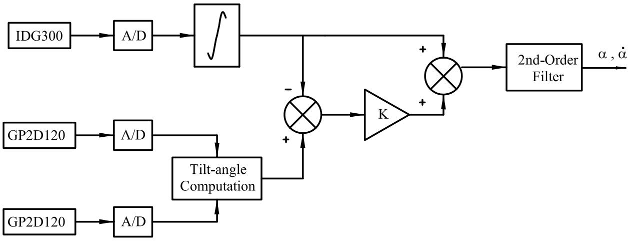


Fig. 7. Fusion of the gyroscope with the distance measuring sensors

It is noteworthy that the adopted fusion scheme is simple and does involve only one filter at the end, rendering it easy and quick for implementation.

Finally, the entities needed for the controller are obtained as follows:

$$\xi_1 = \mathbf{u}_3^T\mathbf{k} = -\sin\alpha, \quad \mathbf{v}_3^T\mathbf{k} = \cos\alpha \ \text{ and } \ \omega_{3l} = \dot{\alpha}$$

5 Experimental Results

The first experiment consists on maintaining InPeRo balanced with zero heading and steering speeds references while applying disturbances. Two types of disturbances were applied: (i) dynamical change of the centre of mass of the central body (by filling a cup of water live) and (ii) giving disturbances by forcing the robot to move forward and backward. The video can be visualized on youtube using the following link: `http://www.youtube.com/watch?v=Xk8_PGsKZX4`.

Figure 8 shows the heading speed and the tilt-angle. Note that despite disturbances, the speed was maintained at zero and the tilt-angle was capable to redress itself.

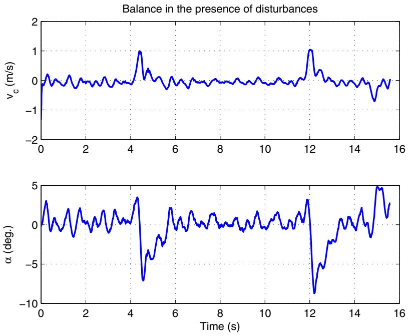


Fig. 8. InPeRo maintaining its central body vertical up while applying disturbances. Both heading and steering speeds are set to zero.

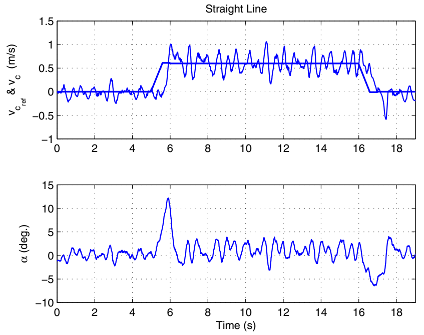


Fig. 9. InPeRo following a straight line: the heading speed follows its reference while the steering speed is maintained to zero

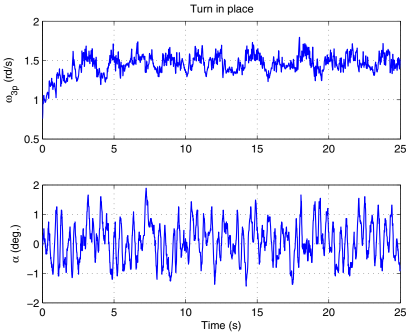


Fig. 10. InPeRo turning in place: the steering speed follows its reference while the heading speed is maintained to zero

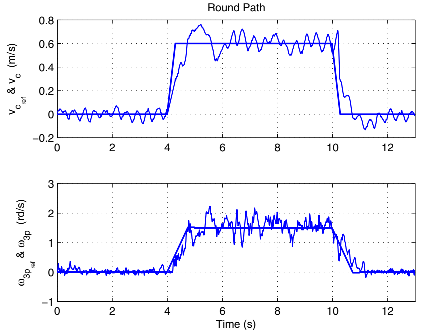


Fig. 11. InPeRo executing a round path: heading and steering speeds follow their references

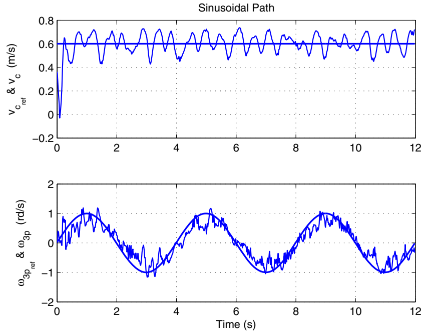


Fig. 12. InPeRo executing a sinusoidal path: heading and steering speeds follow their references

The second experiment consists on following a straight line, i.e., imposing a reference to the heading speed v_c while screwing the steering speed ω_{3p} to zero. Figure 9 shows the results and the corresponding video can be visualized at: `http://www.youtube.com/watch?v=cz6J_gEIepE`.

The third experiment consists on turning in place, i.e., imposing a reference to ω_{3p} while screwing v_c to zero. Figure 10 shows the results.

The fourth experiment consists of executing a round path. Both v_c and ω_{3p} follow their reference as shown in Fig. 11. The video corresponding to this experiment ca be visualized at `http://www.youtube.com/watch?v=15Gk87pPQfs`.

Finally, the last experiment consists on following a sinusoidal path as shown in Fig. 12. The video showing this path is given at: `http://www.youtube.com/watch?v=wy7DfkPlKAE`.

6 Conclusions

In this paper we described the realization of InPeRo, a mobile wheeled pendulum. We recalled the mathematical model of the system and its controller. Then we moved to the construction of the robot including material, actuators and sensors. Finally we gave the experimental results showing the performance of the controlled system. For future work we would like to implement position control and navigation algorithms.

References

1. Kamen, D.L., Ambrogi, R.R., Duggan, R.J., Heinzmann, R.K., Key, B.R., Skoskiewicz, A., Kristal, P.K.: Transportation Vehicles and Methods. US patent 5,971,091 (1999)
2. Grassser, F., D'Arrigo, A., Colombi, S., Rufer, A.C.: JOE: A Mobile, Inverted Pendulum. IEEE Trans. Industrial Electronics 49(1), 107–114 (2002)
3. Salerno, A., Angeles, J.: Design and Implementation of a Quasiholonomic Mobile Robot. In: IEEE Proc. International Conference on Robotics and Automation, Roma, Italy, April 10-14 (2007)
4. Nasrallah, D.S., Angeles, J., Michalska, H.: Modeling of an anti-tilting outdoor mobile robot. In: Proc. 5th Int. Conf. Mutlibody Syst., Nonlinear Dynamics, and Control (ASME), Long Beach, CA (September 2005)
5. Nasrallah, D.S., Michalska, H., Angeles, J.: Controllability and Posture Control of a Wheeled Pendulum Moving on an Inclined Plane. IEEE Trans. Robotics 23(3), 564–577 (2007)
6. Deegan, P., Thibodeau, B.J., Grupen, R.: Designing a Self-Stabilizing Robot for Dynamic Mobile Manipulation. In: Robotics: Science and Systems - Workshop on Manipulation for Human Environments, Philadelphia (August 2006)
7. Anderson, D.P.: `http://www.geology.smu.edu/~dpa-www/robo/nbot/` (last update: March 1, 2010)
8. Brisebois, S., Nasrallah, D.S., Saad, M.: Experimental Validation of Velocity Control of an Inverted Mobile Wheeled Pendulum. In: Proc. 5th International Conference on Industrial Automation, Montreal, Canada, June 11-13 (2007)

9. Toyota Motor Corporation, `http://www.toyota.co.jp`
10. ReconRobotics, `http://www.reconrobotics.com`
11. Jeong, S., Takahashi, T.: Wheeled Inverted Pendulum Type Assistant Robot: Design Concept and mobile Control. Journal of Intelligent Service Robotics 1(4), 313–320 (2008)
12. Lee, H., Jung, S.: Control of a Mobile Inverted Pendulum Robot System. In: Proc. International Conference on Control, Automation and Systems, Seoul, Korea, October 14-17 (2008)

Implementation of Usage Role-Based Access Control Approach for Logical Security of Information Systems

Aneta Poniszewska-Maranda and Roksana Rutkowska

Institute of Information Technology, Lodz University of Technology, Poland
anetap@ics.p.lodz.pl

Abstract. As the technology grows rapidly and the new applications and systems are being developed every day, it is crucial to have proper protection. Information is becoming a strategic asset and because it is often of sensitive nature, it ought to be secured. The paper presents how the Usage Role-based Access Control model introduces improvement to the logical security of information systems. The model is presented in the light of currently used and existing access control models and implemented in a form of a simplified ebook store application.

1 Introduction

Rapid development of information systems and applications in today's world brings with it increased computerization of the enterprises and private homes. Data can now be processed faster and analysed in a more abstract way by the new technologies. Problems can be solved faster then ever before. However this also raises the issue of logical security of data contained in the systems. This data can often be of sensitive nature, like personal data of employees. That is why an access control was introduced. Its main goal is to protect the system resources from undesired user access. Many models of access control are currently available and present in information systems, each having their advantages and disadvantages. The problem analysed in the presented paper is how the new concept of Usage Role-based Access Control deals with the issue of the logical security. It will be also compared with other popular models to see what advantages it introduces.

Security is one of the main issues that come up with the information systems. Due to the fact that a lot of sensitive data is stored on the computers, it has become a real target for hackers. It is therefore important to ensure a proper protection of this data and take all the necessary measures to prevent the unwanted access to it. To say that the stored information is safe, it must satisfy three requirements: only users who are authorized can access it when it is needed (*availability*), unauthorized users cannot modify it (*integrity*) and information can be viewed only by users who have the right to do so (*conffdentiality*) [10].

The presented paper shows how the Usage Role-based Access Control approach introduces improvement to the logical security of information systems.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_13, © Springer International Publishing Switzerland 2014

This approach is presented in the light of currently used and existing access control models. The focus is put on the new concept of Usage Role-based Access Control (URBAC). For practical aspect it was implemented in a form of a simplified ebook store, which presents the main elements of the model in a clear and precise way.

This paper presents the engineering aspects in access control of dynamic information systems. The paper is composed as follows: section 2 presents the role concepts and usage concept in aspect of access control, section 3 gives the outline of approach based on these concepts (Usage Role Based Access Control, URBAC). Section 4 deals with the implementation of URBAC model in a form of ebook store application to properly show its capabilities and elements.

2 Access Control Based on Role and Usage Concepts

Proper authentication by means of an access control is vital to ensure only authorized users can access data and block any unwanted guests or attackers that may pretend to be a legitimate user. In this context, the access control can be seen as both an element of security and a sound basis for implementing the further security measures. By means of access control an organization should have, first and foremost, control from unauthorized use of data and resources. Then it can focus on preventing any intrusions from the outside [11].

The main idea of an access control is to restrict and protect an access to some resource and ensure that only those allowed to use it can access it [10]. Resource can be any element of the system, like a file, folder, database or a printer. Apart from controlling an access to a resource it also deals with how and when the resource can be used [11]. As an example, operating system controls access to the files and a certain user may have access to edit a given file, but only during working hours. The aim of access control is to prevent any unwanted or undesired access to resources. What is more, if properly managed, access control also promotes proper information sharing across users and applications [12].

It is very hard to design an access control model that is perfect and applicable to many types of systems and differing needs. Due to this, each of the traditional access control models or their extensions has some limitations. With rapid development of new systems and applications the needs for control of data are constantly changing with the new problems needing to be solved.

Mandatory Access Control provides very strict and rigid control. It highly limits the user's possible actions and doesn't consider any dynamic alterations of underlying policies [13]. As the policies are managed by a central authority, the main benefit is immunity to Trojan Horse attacks and ensuring the system security, regardless of the users actions. However the main downfall of this model is difficulty to implement in the real-world applications and systems, which have to be rewritten in order to adhere to the model's labelling concept. Another disadvantage is a possibility of over-classifying data by the model, which can affect the productivity of users, who cannot access the data they need.

Discretionary Access Control, unlike MAC gives more freedom to the users. It is left to their discretion to specify the access rules for files they are owners of. The main problem arising in this model is no protection from the copy operation. If a user can read another user's file, there is nothing stopping him from copying that file to a file that he owns. Then he can freely share its contents [11]. Furthermore as granting access to the files is left to the discretion of the users rather then the system, the outcome may be access control policies not reflecting the organization's security requirements [12]. Due to this the maintenance of system and the verification of security policies is very difficult. It also makes the system very vulnerable to the Trojan Horse attack.

Role-based Access Control provides a structure of access control tailored to the needs of enterprises. However it also creates a challenge between easy administration and strong security [12]. For the latter, it is better if the roles are more granular and thus multiple roles are assigned to users. On the other hand, for easier administration it is far more convenient to have less roles to manage. What is more, role engineering may also pose a challenge as an access control may not always be compatible with organization's structure [1].

Usage Control model [3–5] was introduced as an answer to the limitations of the above models. As all of them focused on the authorizations done before an access, this model introduced a possibility to check them also during an access. Furthermore it created a concept of obligations and conditions, which were omitted by earlier models. However the idea of usage control focuses mainly on the management of rights and an access to digital objects. It is a very abstract model, that does not provide a clear structure like role-based access control model and does not deal with who defines and modifies the rights that the subjects posses.

These disadvantages and the needs of present information systems caused the creation of unified model that can encompass the use of traditional access control models and allow to define the dynamic rules of access control. Two access control concepts are chosen in our studies to develop the new approach for dynamic information systems: role concept [6] and usage concept [3, 5]. The first one allows to represent the whole system organization in the complete, precise way while the second one allows to describe the usage control with authorizations, obligations, conditions, continuity (ongoing control) and mutability attributes.

3 Approach of Role Based Access Control with Usage Control

Usage Role-based Access Control (URBAC) [9] is based on a role concept from extended Role-Based Access Control and usage concept from Usage Control. It takes best features from both models in order to combine them into an even more efficient model of an access control. It incorporates the control of usage in data access with authorizations, obligations and conditions that can be applied both before and during access. URBAC also uses a complete and precise way to represent the entire system organization with the use of roles and functions. The main elements of the model are presented in figure 1.

Subject can represent users and groups of users, that share the same rights as well as obligations and responsibilities. Session is the interval of time during

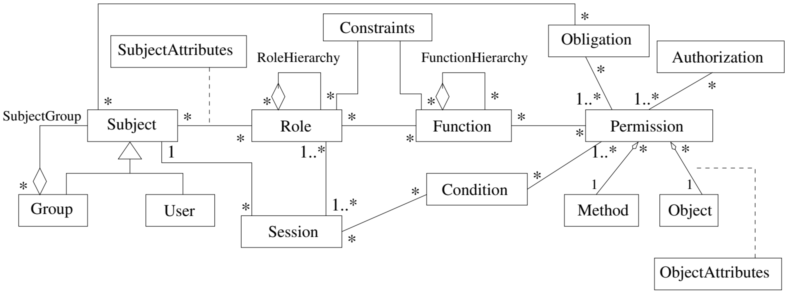


Fig. 1. Meta-model of URBAC approach

which a user is actively logged into the system and may execute the actions in it that require the appropriate rights. User is logged in to the system in a single session, during which the roles can be activated.

A **Role** can be regarded as a reflection of position or job title in an organization, that holds with it the authority as well as responsibilities. It allows to accomplish certain tasks connected with processes in an organization. Users are assigned to them based on their competencies and qualifications. Therefore, role is associated with subjects, where user or group of users can take on different roles, but one role can also be shared among users. This association also contains *Subject Attributes*, like identity or credits, which are additional subject properties that can be considered in usage decision. Role is also associated with a session, which represents the roles that can be activated during a single session. In accordance with Role Based Access Control the role hierarchy can be defined, where inheritance relations are established between the roles.

As each role specifies a possibility to perform specific tasks, it consists of many **functions**, which users may apply. Like with roles, function hierarchy can be defined with inheritance relations between specific functions. Function in turn, can be split to more atomic elements which are operations that are performed on objects. Those are granted by **permissions**. We therefore can view the functions as sets or sequences of permissions, that grant them right to perform the specified methods on a specified objects. To ensure that the model is coherent, each existing permission must be assigned to at least one function.

In the model when permissions are assigned to objects, the specification of constraints is necessary. Those constraints are authorizations, conditions and obligations. Constraint determines that some permission is valid only for a part of the object instances. We can denote a permission as a function: *p(o, m, Cst)* where*o* represents an object, *m* a method that can be executed on the object and *Cst* is a set of constraints that determine this permission. Taking into consideration a concept of authorization, obligation and condition, the set of constraints can take the following form $Cst = \{A, B, C\}$ and the permission can be presented as a function $p(o, m, \{A, B, C\})$. According to this, the permission is given to all instances of the object class except the contrary specification.

The **constraints** are defined in accordance with Usage Access Control model. **Authorizations** are logical predicates attached to a permission and determine permission validity. They depend on the access rules, object attributes and subject attributes. **Obligations** are functional predicates which verify mandatory requirements that a user has to perform before or during access. In the model they are defined for permissions, but also concern the subjects. **Conditions**, in turn evaluate the current status of the system as well as environment to check if relevant requirements are satisfied. They are defied for permissions, but concern also the session. Conditions are subject and object independent. All these three elements can be either checked before access request (*pre*) or can be checked continuously or periodically during subject's access to the object (*ongoing*). Furthermore the concept of constraint can be also defined for main elements of the model (user, group, subject, session, role, function, permission, object and method) as well as for relationships between the elements.

Objects are entities that can be accessed by users. They have a direct relationship with permissions that can be further described with the use of Object Attributes. Those represent additional object attributes specific to the relation, like for instance the security labels or ownerships. Attributes of both subjects and objects can be mutable, which means they can be updated by the system as consequences of subject usage on objects. Attributes can also be immutable and cannot be changed by the system, but only at administrator's discretion.

4 Implementation of Usage Role-Based Access Control Approach

Usage Role-based Access Control approach was implemented for the practical aspects in a form of simplified ebook store application that presents the main elements of the model in a clear and precise way. Based on the created application the URBAC approach was compared to the other access control models.

4.1 Application of Ebook Store

The concept used to illustrate the Usage Role-based Access Control is an online ebook store. It gives many possibilities for the role definitions as well as the functions to be performed by the users. This concept allows to represent the access control in an extensive way. It also deals with an access to digital objects, which is a large part of Usage Control, that is a part of URBAC. The main capabilities in the store are:

- sell and purchase of the ebook files,
- downloading the ebook files,
- purchase of the credit,
- providing the feedback,
- editing the ebook items,
- featuring the ebooks,
- upgrading the user roles.

Terms of Service were introduced that the users should agree to be able to perform the most functions in the store including purchase. The terms encompass the information regarding the user's accounts, security rules for maintaining and using these accounts as well as the usage rules of the store. If a user does not abide by these terms, his account may be terminated by the administrator. Additionally, as the store itself does not deal with copyright at the moment, for selling ebooks users will have to confirm the copyright agreement.

There are different types of roles available for the users - **Regular**, **Premium** and **Seller**. There is also a special role of the **Administrator**, who has access to all the application models and data to define the new elements of the access control in accordance with the Usage Role-based Access Control model directly without the need to modify the application's code.

There are many possible functions for an online store. The basic ones will be registering in the service as well as logging in and out. The registered users will have the possibility to obtain different roles: *Regular*, *Premium* and *Seller*. Regular User functions are: register, login, logout, browse the available ebooks, view information about a particular ebook file, view a user's profile, purchase an ebook, search for an ebook, add an ebook into the cart, view the user's cart, payment with credits, purchase of the credits, download the bought ebook file with 30 seconds wait, upgrade the account. Premium User functions are: all the functions of the Regular User, no wait time for the downloads, giving feedback to other users, more payment options (credits, credit card or paypal). Seller User functions are: all the functions of the Premium User, sell an ebook item, edit an ebook item, feature an ebook item, remove an ebook item. Administrator functions are: all the functions available to the users, modify and create the data contained in the application, including the elements of URBAC.

4.2 Proposed Implementation of URBAC Model

Creation and modification of the elements of URBAC model are realized by the security administrator with the use of an administrator's panel. The administrator is able to create the security policy rules dynamically at the level of the application and these policy rules are applied to the ebook store application immediately. The framework used for the creation of the application allowed to generate an attractive interface for the administrator. Through this interface the administrator can see and adjust the lists of created elements and the connections between them. The elements of the URBAC approach are created according to schema presented in figure 1.

As described above there are four main types of available roles: Regular, Premium and Seller users and a special role of administrator. Administrator is able to modify the attributes of these roles, like their upgrade price or decide to create a new type of role available in the application. Each role consists of a collection of functions. Therefore the administrator is able to assign different functions to various roles. Examples of the main available functions are: buy an ebook, download an ebook, edit an ebook, register, remove an ebook, upgrade an account, sell an ebook, get credits, give feedback, feature an ebook.

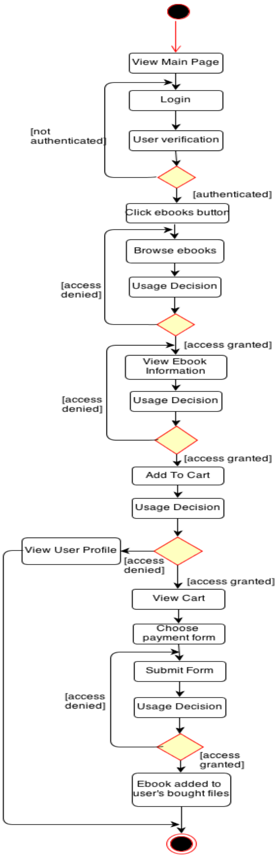


Fig. 2. Activity diagram for "buy" function

For instance the function to *edit an ebook* is assigned only to the role *Seller*, while the function *Buy an Ebook* is assigned to all the roles available in the application. To perform a function a user has to make a sequence of activities. For instance to buy an ebook, as shown in figure 2, the user has to first log in, view the available ebooks, choose one he wants to buy, add it to his cart, and so on. Some **permissions** are assigned to each of these activities. For instance in case of the *buy function* the permissions are: view ebook list, view ebook information, add an item to the user's cart, view the user's cart, purchase the item.

When a permission is evaluated it consists of *authorizations*, *obligations* and *conditions*. All three have to be satisfied to give a user access and let him proceed to next step in the function. Examples of *authorizations* for the *purchase of the ebook* permission, that the administrator can define, presented in form of questions, can be as follows:

- Does a user own the ebook object?
- Does the user have enough credits to purchase the ebook?
- Has the user already bought the item?
- Does the user have any downloads of the ebook left?

Obligations for instance, can consider if the user has agreed to the terms of service and the copyright agreement. If he has not agreed his activities in the application will be limited. *Conditions*, in turn, can focus on allowing the users to execute the functions only during the business hours and setting a limit on the number of ebooks that can be featured in the store.

The object of *Authorization* class evaluates a logical statement if a user can be granted access. This statement can take into consideration the following types of comparisons:

- compare a user's attribute with the fixed values,
- compare an ebook's attribute with the fixed values,
- compare ebook's and user's attributes with each other,
- compare ebook's attribute with an attribute of another object retrieved from the database,
- compare user's attribute with an attribute of another object retrieved from the database,
- find a single ebook object in one of user's lists: ebook downloads, owned ebooks, cart items.

For instance an *Authorization* object can check if a user has sufficient funds to purchase an ebook object. It is checked if the user's attribute *credits* value (integer data type) is greater than the value of ebook's attribute *price*.

The object of *Obligation* class evaluates if a user satisfies the mandatory requirements (actions) he has to perform before access. The *Obligation* object focuses on comparing the user's attributes with different data types. These data types have fixed values. For example an *Obligation* object can check if a user has agreed to the terms of service. To do this, the user's attribute *termsAgreedOn* is compared with boolean "true" value.

The objects of the *Condition* class are used to evaluate the requirements that depend on the environment and on the system. These requirements can be therefore based on different aspects:

- particular value taken from the framework's cache,
- time-specific evaluation,
- particular system variable retrieved from database.

For instance in case of time-specific evaluation, the business hours can be set with integer data type. In order to do that, the fixed values must be specified - the low value is set to 8 and the high value is set to 16. When the condition is evaluated the current time is checked and compared against the fixed values. If this condition is satisfied the user can perform his activities between 8 and 16 o'clock. If he wants to gain an access outside of these hours, he will get an error string message.

5 Conclusion and Discussion

Rapid development of new technologies brings with it a need for the new security solutions. Usage Role-based Access Control introduces a new approach to the logical security of information systems. It is a very flexible model that can be tailored to various needs.

The most important advantages provided by URBAC model include:

- clear structure, with well-described elements that are logically linked,
- fit for enterprise needs with roles assigned according to the job positions and functions that reflect the employee's tasks and responsibilities,
- security administrator can create the policies that reflect an organization's security policies,
- model takes into account that attributes can change,
- access permissions are evaluated dynamically at the time of access request and therefore may depend on changing attributes,
- increased flexibility of permissions, which are based not only on the authorizations, but also on the concept of obligations that a user has to perform for an access and the concept of conditions that reflects a state of the system and its environment,
- user's permissions are evaluated not only before but also during an access, thanks to the authorizations, obligations, conditions and subject and object attributes,
- clear description of the model provides a great basis for the creation of administration panel to manage the security policies dynamically without system's downtime.

Compared to the other models the URBAC approach introduces much more elements and covers broader needs of the security. MAC creates very rigid control enforced by the system, while in URBAC the control is focused more on dynamically changing attributes, with permissions computed at an access request. DAC introduces more control at the discretion of users, which may lead to many discrepancies with the organization's policies. In URBAC it is the administrator who defines and manages the policies, tailoring them to the needs of an enterprise. RBAC, in turn provides a clear structure. It is easily adapted for the organizations with the roles concept. However it lacks more flexible permissions and their evaluation during an access, which can be found in UCON. This is why Usage Role-based access control approach incorporates the best features from both RBAC and UCON to create an even more efficient and flexible access control.

The application created for implementation of URBAC approach was an ebook store website. It offered a wide range of functionalities. The main roles offered to the users are Regular, Premium and Seller. Each encompasses a different range of functionalities. There is also a special role of an administrator that allows to dynamically create new rules of security policy for the ebook store application. This is done through an administration panel. It provides an easy interface to manage the URBAC elements and other application data. Access to each functionality in the store is guarded by the access rules and a user is informed if any are not satisfied through appropriate messages displayed in the store's view.

The URBAC approach creates a sound and effective base for the proper protection of data. Its implementation confirms that URBAC can be distinguished as very well solution to the needs of modern enterprises.

References

1. Ferraiolo, D., Sandhu, R.S., Gavrila, S., Kuhn, D.R., Chandramouli, R.: Proposed NIST Role-Based Access control. ACM TISSEC (2001)
2. Park, J., Zhang, X., Sandhu, R.: Attribute Mutability in Usage Control. In: 18th IFIP WG 11.3 Working Conference on Data and Applications Security (2004)
3. Lazouski, A., Martinelli, F., Mori, P.: Usage control in computer security: A survey. Computer Science Review 4(2), 81–99 (2010)
4. Pretschner, A., Hilty, M., Basin, D.: Distributed usage control. Communications of the ACM 49(9) (September 2006)
5. Zhang, X., Parisi-Presicce, F., Sandhu, R., Park, J.: Formal Model and Policy Specification of Usage Control. ACM TISSEC 8(4), 351–387 (2005)
6. Poniszewska-Maranda, A.: Conception Approach of Access Control in Heterogeneous Information Systems using UML. Journal of Telecommunication Systems 45(2-3), 177–190 (2010)
7. Strembeck, M., Neumann, G.: An Integrated Approach to Engineer and Enforce Context Constraints in RBAC Environments. ACM TISSEC 7(3) (2004)
8. Bertino, E., Ferrari, E., Atluri, V.: The Specification and Enforcement of Authorization Constraints in Workflow Management Systems. ACM TISSEC 2(1)
9. Poniszewska-Maranda, A.: Modeling and design of role engineering in development of access control for dynamic information systems. Bulletin of the Polish Academy of Sciences, Technical Science (accepted, 2013)
10. Kim, D., Solomon, M.: Fundamentals of Information Systems Security. Jones & Bartlett Learning (2012)
11. Ferraiolo, D.F., Kuhn, D.R., Chandramouli, R.: Role-Based Access Control, 2nd edn. Artech House (2007)
12. Hu, V.C., Ferraiolo, D.F., Kuhn, D.R.: Assessment of Access Control Systems, Interagency Report 7316, NIST (2006)
13. Stewart, J.M., Chapple, M., Gibson, D.: CISSP: Certified Information Systems Security Professional Study Guide, 6th edn. John Wiley & Sons (2012)

Analytical Possibilities of SAP HANA – On the Example of Energy Consumption Forecasting

Tomasz Rudny, Monika Kaczmarek, and Witold Abramowicz

Department of Information Systems, Faculty of Informatics and Electronic Economy,
Poznan University of Economics
tomasz.rudny@ue.poznan.pl

Abstract. The vast amount of data that organizations should gather, store and process, entails a set of new requirements towards the analytical solutions used by organizations. These requirements have become drivers for the development of the in-memory computing paradigm, which enables the creation of applications running advanced queries and performing complex transactions on very large sets of data in a much faster and scalable way than the traditional solutions. The main aim of our work is to examine the analytical possibilities of the in-memory computing solution, on the example of SAP HANA, and their possible applications. In order to do that we apply SAP HANA and its components to the challenge of forecasting of the energy demand in the energy sector. In order to examine the analytical possibilities of SAP HANA, a number of experiments were conducted. Their results are described in this paper.

Keywords: in-memory computing, SAP HANA, energy demand forecasting.

1 Introduction and Motivation

There is a vast amount of data that organizations should collect, store and analyze for the needs of a decision making process. Data is entered in batches or record by record, using multiple channels [1]. Roughly the data may be divided into four categories: transactional data from daily operations, stream data (e.g., coming from various sensors), structured and unstructured data published, e.g., on the Internet. The results of the analysis conducted on all of the mentioned categories of data should be used in order to streamline the already running business processes or should influence the definition of the strategic objectives of an organization. Thus, on the one hand, business users need a fast and reliable access to information and various analyses, in order to quickly respond to the changes within or outside of the organization. However, on the other hand, organizations must consider the presumably high costs associated with the purchase and maintenance of information technology-based solutions to store and process a vast amount of data [2].

These challenges have become drivers for the development of the in-memory computing paradigm, which enables the creation of applications running advanced queries and performing complex transactions on very large sets of data

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_14, © Springer International Publishing Switzerland 2014

“at least one order of magnitude faster - and in a more scalable way“ [3], in comparison to the traditional architectures. This is achieved by storing data in the main Dynamic Random Access Memory (DRAM) instead of on electromagnetic disks.

In-memory computing on the one hand, allows to carry out batch processes in minutes rather than in hours, enable the fast delivery of process to various stakeholders, supports much more advanced data mining and correlations analysis of millions of records in seconds. On the other hand, the decreasing prices of semiconductor technologies together with rapidly developing application infrastructure technologies, are contributing to the increasing adoption of the in-memory computing solutions, due to their more competitive prices [3]. As a result, in-memory computing provides an opportunity to change the way in which organizations fulfil their business requirements [2].

The main aim of our work is to examine the analytical possibilities of the in-memory computing solution, on the example of SAP HANA, and their possible applications. In order to do that we apply SAP HANA and its components to the challenge of forecasting of the energy demand in the energy sector. Forecasting energy demand in real time as well as creation of models featuring a low prediction error is extremely difficult [4]. Therefore, such a scenario requires a solution that supports fast operations on a very large number of data and is equipped with adequate analytical modules. These requirements correspond to SAP HANA.

Thus, our aim is to examine, based on the energy sector the following use case: (1) the built-in analytical tools provided by SAP HANA and the ease of their usage, (2) the efficiency of the process of processing of large set of data and its scalability. In order to reach the above mentioned goals, a number of experiments were conducted on the SAP HANA instance, and their results are presented in this paper.

The paper is structured as follows: First the SAP HANA is briefly presented as an example of in-memory computing solution. Then, the specific aspects and challenges connected to the forecasting of the energy demand are given. Finally, the conducted experiments and their results are presented. The paper concludes with final remarks and an outlook on the possible applications of in-memory computing.

2 SAP HANA

In-memory database is a database management system with its storage employed in the RAM instead of hard drives based memory. Because hard drives are a factor or two slower, this offers a huge performance improvement. Such approach eliminates the time needed to seek sectors on disks. In-memory databases are usually equipped with a multi-processor and multi-core processing units, which makes them very efficient.

SAP HANA is an example of a modern In-memory database and was the solution analysed within the research carried out. SAP HANA Database (SAP HANA DB) was designed to provide [1,2,5]:

- a main-memory centric data management platform that should support SQL as well as additional interaction model for the needs of SAP applications,
- transactional behavior for the needs of interactive business applications,
- adequate parallelization.

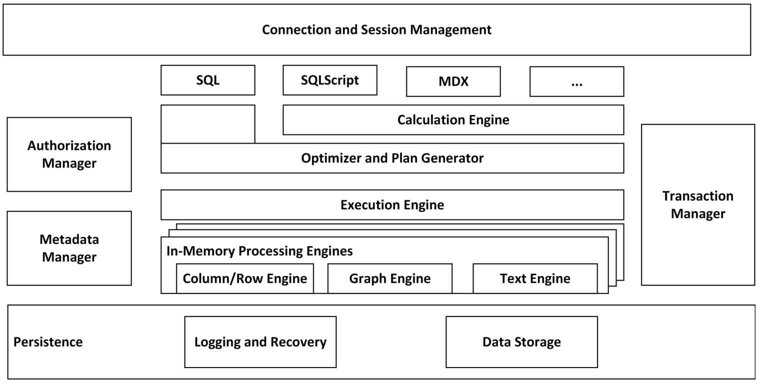


Fig. 1. Overview of SAP HANA DB [1]

The architecture of the SAP HANA DB is shown in Fig. 1. The main component of SAP HANA is the calculation engine, also known as the SAP In-memory Computing/Processing Engine [1]. It keeps data in the main memory (RAM) for as long as there is a space available, which can be in TB. The calculation engine operates together with other processing engines and communicates with outside applications via JDBC, ODBC, ODBO, SQL DBC and other protocols [6]. This provides access via SQL, MDX and BISC languages for many business applications like SAP Business Objects BI clients.

Data resides in memory in tables, which are in column or row layout. The column order tables are optimized structures for in-memory processing and provide high efficiency of the solution [1, 2, 5].

2.1 R Integration

One way to integrate analytical processing in SAP HANA based systems is to use R. R is a popular open source software framework for statistical analysis [7]. There are more than 3000 packages and libraries available on the Internet. Moreover, R is a known language and can be positioned as the popular standard among scientists and analysts. Therefore, it is highly beneficial that one can use R with SAP HANA.

The R integration is based upon setting up a server for R called Rserve. The communication is achieved by means of TCP/IP protocol, however, it is fast because both the SAP HANA instance and the Rserve are usually located in the same local network or even in the same physical machine. The R code is put inside the SQLScript, which is a SAP HANA's extension of SQL. SQLScript allows for definition of arbitrary processing logic and flow control within the database procedures. The R code, thus embedded inside SQLScript procedures, is then submitted as a part of a query. The calculation engine of the SAP HANA receives the code as a string argument, transfers intermediate database tables directly into the vector oriented data structures of R [8]. The idea is to improve performance over standard SQL interfaces, which are tuple-based and therefore, require an additional data copy on the R side. Then, the calculation engine sends a request through the Rserve mechanism to create a dedicated R process on the R host. When the R process finishes the execution of the R code, then the results - necessarily as a data frame - are sent back to the calculation engine and there are converted back into the column-oriented data structures of SAP HANA. This conversion is efficient [8].

However, the above-mentioned process makes it more difficult for the analyst or programmer to write programs. This is because, there are several limitations put both on the input and output parameters of the R-code based procedures. The other limitations of using R in SAP HANA currently are [8]:

- Only table types are supported as parameters in SQLScript procedures of language RLANG, so scalar types must be also passed as table types.
- None of the flexible R data types as lists or matrices can be passed from R to SAP HANA, unless converted to data frames.
- All embedded R functions must have at least one result in the form of a data frame.
- Variable names in the procedures may not contain upper-case letters.
- Factor columns can only be retrieved as character vectors.

Despite those limitations R remains a powerful analytical tool accessible in SAP HANA based systems. Its wide range of packages provides a flexible framework for analytical processing.

2.2 Predictive Analysis Library

Apart from R integration SAP HANA Suite offers its own analytical toolbox - Predictive Analysis Library (PAL) [9]. PAL functionality is accessible by means of SQLScript. The idea behind PAL is to eliminate the overhead of data transfer from the database to the analytical application, thus increasing the overall performance of the system.

PAL provides a wide range of analytical functions that can be invoked directly from SQLScript, which means they can be run on the database. The key benefit of PAL over R is that no cumbersome conversions between data types are required. Using a PAL function can be done with 3 simple steps:

1. Definition of the so-called AFL_WRAPPER_GENERATOR procedure (done only once).

2. Definition of the user procedure in SQLScript that wraps the PAL function.
3. Actual call of the procedure.

However, the functionality set of PAL is limited in comparison to thousands of packages available in R. Currently, PAL provides functions for [9]: Clustering, Classification, Association analysis, Time Series forecasting, Other algorithms - e.g., sampling, binning, range tests.

The functions used in our research are those of Time Series group. Here PAL provides only 3 models at the moment, although based on SAP claims this can change for better [9].

3 Forecasting of the Energy Demand

The energy market sector is facing many challenges – e.g., usage of renewable energy sources, application of smart metering, emergence of new players, as well as accurate forecasting of a short- and long-term value of energy demand and energy production from different sources.

Erroneous forecasts of energy demand entail costs, resulting, among others, from the need to purchase the additional energy capacity at the energy balancing market for a significantly higher price. However, forecasting energy demand in real time as well as creating models featuring a low prediction error is extremely difficult. Although in practice there are many approaches applied, even an intimate knowledge of computational techniques and stochastic methods does not allow to achieve the desired results without the support provided by the advanced and modern IT technologies.

Taking into account a large number of influence factors and their uncertainty, ”it is not possible to design an exact physical model for the energy demand” [10]. Therefore, the energy demand is calculated using statistical models (e.g., regression models, probabilistic models [11, 12]), artificial intelligence tools [13] or hybrid approaches [14, 15] trying to capture and describe the influence of climate factors, operating conditions [10] and other variables, on the energy consumption.

Most of the existing approaches to energy demand prognosis focuses on characterising aggregated electricity system demand load profile. However, the energy load forecasting can be done more accurately, if forecasts are calculated for all customers separately and then combined via the bottom-up strategy to produce the total load forecast [12, 16]. The reason for it is that the characteristics of aggregated and individual profiles are different, e.g., they have different shapes - the individual profiles show peaks in the morning and evenings and their shapes vary for different days of the week and parts of the year. In fact, the individual electricity load profile is influenced by the number of factors, which may be divided in the following categories [12]: (1) electricity demand variations between customers, (2) seasonal electricity demand effects, (3) intra-daily variations, and (4) diurnal variations in electricity demand. The forecasting on an individual level can lead to substantial savings for companies and also for the environment, up till now however, it was computationally difficult, as for a typical energy seller a number of residential customers exceeds hundreds of thousands. The availability of the in-memory computing solution can change that.

4 Experiments

Within the conducted experiments, we focus on time series approach in order to predict the short-term energy demand value. We apply the time series method at an individual dwelling level (similarly to [12]) as well as at an aggregated level and take advantage of SAP HANA and its components (see Fig. 2).

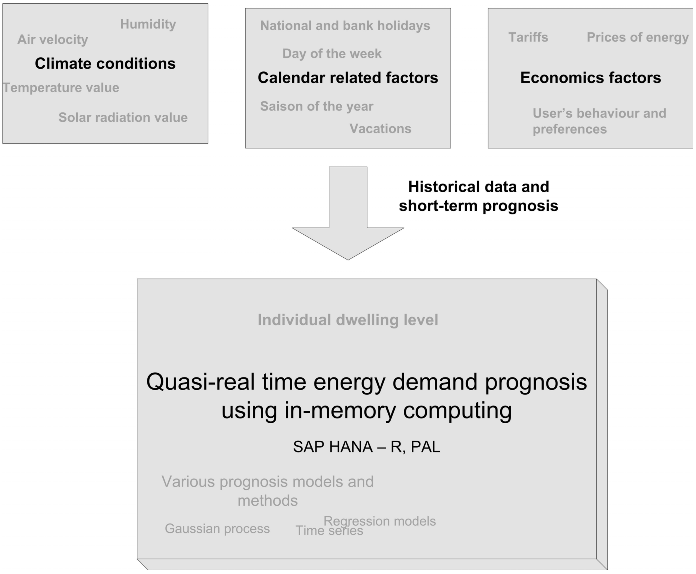


Fig. 2. Short-term energy demand forecasting on the individual level using SAP HANA

The conducted experiments with SAP HANA were two-fold. On the one hand, the possibility to apply R and PAL analysis functionalities to generate forecast of energy demand was to be tested. On the other hand, the computational capabilities of SAP HANA were also in the centre of our attention.

4.1 Experiments with PAL

The models implemented in PAL are: single, double and tipple exponential smoothing. Exponential smoothing is often used in predicting stock exchange values as it reasonably well detects trends in the data, as well as models regular changes. However, it often performs poorly with highly seasonal data that varies greatly.

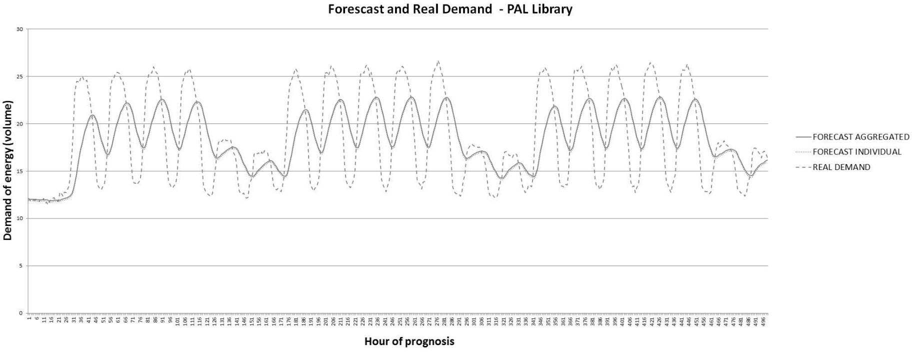


Fig. 3. Forecast and Real Demand

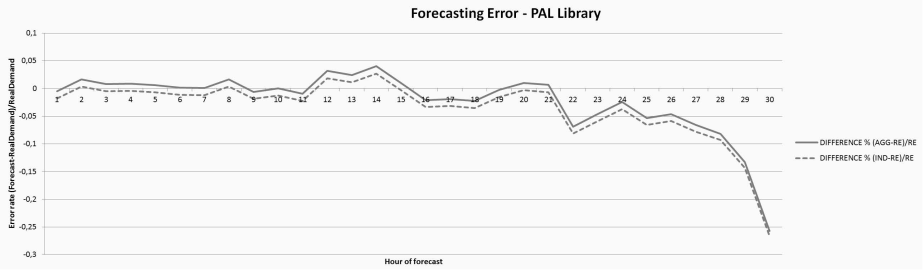


Fig. 4. Error of Prognosis

So for the purpose of energy demand forecasting exponential smoothing models are not the best option (see Fig. 3 and Fig. 4). However, in this paper we focus on other aspects of the problem, trying to analyze SAP HANA capabilities. The forecasts accuracy, although critical in real applications, was not the main issue here. However, when new models are introduced into PAL, then it should also significantly improve.

4.2 Experiments with R

We were able to generate forecasts using various models based on 180 days historical data with a 14 days horizon. The models tested included ARIMA and Holt-Winters exponential smoothing. The results show that SAP HANA can be efficiently used for this purpose.

The runtime for calculating summary forecast using the bottom-up approach varies around 600 *ms* as shown on the histogram (see Fig. 5). It is not only possible calculate summary forecast in the quasi-real time, but also to run different models and choose the best fitted one for each customer.

The code for forecasting can be nicely and neatly written in R as shown on the example of Holt-Winters exponential smoothing:

```
CREATE PROCEDURE FS_STEP(IN x "input2",
IN temp "temp", OUT y "totals_forecast")
LANGUAGE RLANG AS
BEGIN
    ts1 <- ts(x$value, frequency=24)
    m <- HoltWinters(ts1)
    f <- predict(m, n.ahead=14*24)

    forecast <- f[1:14*24]

    y <- as.data.frame(cbind(temp, forecast))
END;
```

4.3 Efficiency of Computations

As stated at the outset our aim was to analyze possibilities of applying SAP HANA to complex analytical processing on the example of energy demand forecasting. One of the key requirements for any tool in this domain is computational efficiency. We implemented and run a series of benchmarking experiments to see if SAP HANA meets the challenge.

In the experiments we used original demand data obtained from one of the major electrical sellers in Poland. The data were anonymized to protect the customers privacy. The data spanned 8 months from January till August. Measurements of electricity consumption were taken every hour in the given period.

Figure 5 presents the histogram of runs of R code for the individual time series forecasting algorithm. This algorithm first computes forecasts at the individual level, then summarizes them to obtain the final forecast. As can be seen the results are acceptable for business applications and the run-times are lower than for any desktop based statistical systems. For example the same data processing using a desktop PC with 4 GB RAM and double-core processor took approximately 3.2 seconds using the R package and 2.75 seconds using SAS Forecast Studio. Of course, comparisons between desktop based systems and in-memory

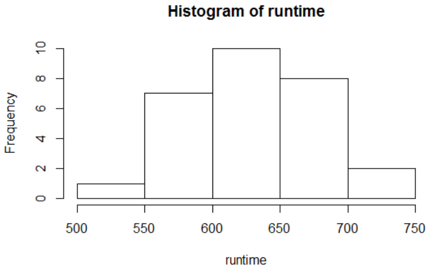


Fig. 5. Runtime histogram of R code for demand forecasting on individual level

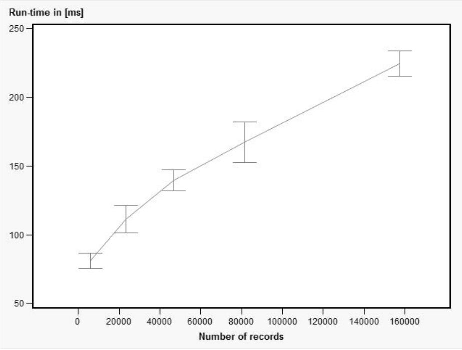


Fig. 6. Scaling of PAL forecasting procedure

computing enabled machines are by nature difficult, however, it shows that SAP HANA can be successfully used for the presented problem.

As hoped, the SAP HANA Predictive Analysis Library (PAL) offers better performance. Especially promising is the scalability of the PAL code as shown in Fig. 6.

5 Conclusions

In this paper we analysed the capabilities of SAP HANA In-memory database for analytical processing. We showed that SAP HANA can be successfully used for complex analyses like time series forecasting, even when applying non-trivial approach based on the individual users energy demand forecasting. Moreover, SAP HANA outperforms classical solutions due to its parallelism and in-memory processing. This shows that classical analytical systems may soon be completely substituted by modern in-memory solutions. The progress in in-memory software development should provide further ease for programmers and analysts removing all interface inconveniences that are inherent in newly released systems.

Acknowledgements. The work described in this paper was carried out within the project *Forecasting of Energy Load Demand and Energy Production from Renewable Sources using In-Memory Computing*, run under the Hasso Plattner Institute (HPI) Future SOC Lab, which provides an access to the SAP HANA system.

References

1. Farber, F., May, N., Lehner, W., Grosse, P., Muller, I., Rauhe, H., Dees, J.: The SAP HANA database – an architecture overview. IEEE Data Eng. Bull., 28–33 (2012)
2. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)

3. Gartner, P.R.: Gartner says in-memory computing is racing towards mainstream adoption (April 3, 2013), http://www.gartner.com/newsroom/id/2405315
4. Weron, R.: Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. John Wiley and Sons Ltd. (2006)
5. Färber, F., Cha, S.K., Primsch, J., Bornhövd, C., Sigg, S., Lehner, W.: Sap hana database: data management for modern business applications. SIGMOD Rec. 40(4), 45–51 (2012)
6. Bernard, M.: SAP High-Performance Analytic Appliance 1.0 (SAP HANA) - A First Look at the System Architecture (2011)
7. Aragon, Y.: Séries temporelles avec R. Méthodes et cas, 1st edn. Springer, Collection Pratique R (2011)
8. SAP: SAP HANA R Integration Guide (2013)
9. SAP: SAP HANA Predictive Analysis Library (PAL) Reference (2012)
10. Schellong, W.: Energy Demand Analysis and Forecast. In: Energy Management Systems, pp. 101–120. InTech (2011)
11. Wang, J., Ma, X., Wu, J., Dong, Y.: Optimization models based on gm (1) and seasonal fluctuation for electricity demand forecasting. International Journal of Electrical Power & Energy Systems 43(1), 109–117 (2012)
12. McLoughlin, F., Duffy, A., Conlon, M.: Evaluation of time series techniques to characterise domestic electricity demand. Energy 50, 120–130 (2013)
13. Zadeh, S., Masoumi, A.: Modeling residential electricity demand using neural network and econometrics approaches. In: 2010 40th International Conference on Computers and Industrial Engineering (CIE), pp. 1–6 (July 2010)
14. Kiran, M.S., Ozceylan, E., Gunduz, M., Paksoy, T.: A novel hybrid approach based on particle swarm optimization and ant colony algorithm to forecast energy demand of turkey. Energy Conversion and Management 53(1), 75–83 (2012)
15. Shakouri, G.H., Nadimi, R., Ghaderi, F.: A hybrid tsk-fr model to study short-term variations of the electricity demand versus the temperature changes. Expert Systems with Applications 36(2, pt. 1), 1765–1772 (2009)
16. Charytoniuk, W., Chen, M.S.S., Kotas, P., Van Olinda, P.: Demand forecasting in power distribution systems using nonparametric probability density estimation. IEEE Transactions on Power Systems 14(4), 1200–1206 (1999)

An Efficient Algorithm for Sequential Pattern Mining with Privacy Preservation

Marcin Gorawski[1,2] and Pawel Jureczek[1]

[1] Silesian University of Technology,
Institute of Computer Science,
Akademicka 16, 44-100 Gliwice Poland
{Marcin.Gorawski,Pawel.Jureczek}@polsl.pl
[2] Wroclaw University of Technology,
Institute of Computer Science,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
Marcin.Gorawski@pwr.wroc.pl

Abstract. This paper presents an index-based algorithm named SSAPP for exploring frequent sequential patterns in a distributed environment with privacy preservation. The SSAPP algorithm uses an equivalent form of a sequential pattern to reduce the number of cryptographic operations, such as decryption and encryption. In order to improve the efficiency of sequential pattern mining, the SSAPP algorithm keeps track of patterns in a tree data structure called SS-Tree. This tree is used to compress and represent sequences from a sequence database. Moreover, a SS-Tree allows one to obtain frequent sequential patterns without generation of candidate sequences. The conducted experiments show the effectiveness of the proposed approach. The SSAPP algorithm greatly reduces the number of cryptographic operations and it has good scalability.

1 Introduction

Centralized data mining gives more information compared to an environment where every individual explores its data independently. In the centralized data mining model all the data are sent to or are stored at a central site which is trusted by all the individual sites. Another way is to not share data and run data mining at each site separately and then combine the results. However, this simple approach provides less accurate global results. For example, global results may not include cross-site correlations when some data are partitioned over multiple sources. Due to security concerns, a trusted site can by replaced with a secure multi-party computation protocol.

In this paper we propose an efficient secure multi-party protocol under semi-honest model for sequential pattern mining [1] - [5]. The semi-honest model is a model where each site follows honestly the protocol but may try to deduce information about the other site's private data by analyzing information received as a result of performing the protocol.

A important part of our secure multi-party protocol is a new algorithm named SSAPP (Selective Search Algorithm with Privacy Preservation) which extends CDKSU (Secure Union with Common Decrypting Key) algorithm. The SSAPP algorithm provides a better performance than the AprioriAll algorithm. A significant performance

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_15, © Springer International Publishing Switzerland 2014

gain is achieved by using a short form of sequential patterns (reduces the number of cryptographic operations) and a tree data structure (i.a., improves performance of pattern generation). To the best of our knowledge, this is the first work addressing these research issues in the context of privacy preservation.

Notice that this paper continues the work started in [6] (see also [7]), where a basic algorithm for continuous pattern mining is presented.

The remainder of the paper is organized as follows. Section 2 presents a multi-party computation protocol. Section 3 gives a general overview of the new approach. In Section 4, we discuss the impact of different parameters on the runtime of the SSAPP algorithm. Section 5 summarizes our work.

2 Cryptosystem

In this paper an efficient secure multi-party protocol under semi-honest model for sequential pattern mining is proposed. In our protocol we assumes that different sites (data owners) do not want to disclose their private data (which might be sensitive or valuable), but still they are willing to mine the union of their databases without using a central site having access to all the data. Since during the computation one could deduce something about parties' private data (e.g., the identity of an owner of a particular sequence may be revealed), all exchanged sequences are encrypted by all parties and since only the full encryption allows for decryption using the common key, it is not possible to associate sequences to their owners.

In order to encrypt and decrypt data, our protocol uses a commutative cipher. The commutative cipher is a cipher in which the order of encryption and decryption is interchangeable. For example, let E_i and D_i be, respectively, the encryption and the decryption by site i and let m be a message (plaintext). If m is encrypted using E_1 and E_2 (i.e., $E_2(E_1(m))$), then to get the fully decrypted message m, the ciphertext can be decrypted using firstly D_2 and later D_1, or D_1 and later D_2.

Notice that our commutative cipher works on fixed-length units called blocks. Therefore, if we want to encrypt a very long plaintext then it is divided into separate blocks and these blocks are encrypted separately. Obviously, the longer a message is, the more time is taken to encrypt the whole message. A similar situation occurs when a ciphertext is decrypted. Therefore, the SSAPP algorithm limits the length of sequences broadcasted to participating sites. By reducing the length of sequences, the number of data blocks to encrypt and decrypt decreases. This in turn leads to better overall performance.

Moreover, as mentioned earlier, each site encrypts data blocks using its private key and only one site decrypts fully encrypted data to obtain a plaintext. A key idea behind this approach is based on the observation that the more sites are involved in secure pattern mining, the more time is taken to complete a task.

2.1 CDKSU Algorithm

The SSAPP algorithm extends the CDKSU algorithm [5]. The CDKSU algorithm is a distributed association rules mining algorithm for horizontally partitioned data, which preserves data privacy in semi-honest model. In order to reduce the number of expensive decryption operations, the CDKSU algorithm uses a common decryption key (CDK) to decrypt every ciphertext only once.

The CDKSU algorithm assumes that a transaction database *DB* is partitioned horizontally over n sites S_1, S_2, ..., S_n in such a way that DB_i resides at site S_i. Every transaction in DB_i is a set of items. An itemset that contains *k* items is called *k*-itemset. The support of a *k*-itemset is defined as the ratio of the number of transactions containing it to the total number of transactions in DB. An itemset is considered to be frequent if its support is no less than a user defined support threshold. Moreover, the locally frequent itemset at site S_i is an itemset that is frequent with respect to DB_i. Notice that a locally frequent itemset may not be globally frequent, i.e., it may not be frequent with respect to *DB*. Furthermore, each site S_i has its own secret key which is not shared with any other site and one site is selected as site *D* which is responsible for decrypting data. The main steps of the CDKSU algorithm are:

1. Each party computes the intersection of the set of locally frequent itemsets and the set of globally frequent itemsets, and then generates candidate itemsets according to the Apriori property [7]. Next, the support of each candidate is counted and finally locally frequent itemsets are obtained.
2. After all locally frequent itemsets are discovered by a party, they are encrypted using the secret key and sent to a next party. When a party receives itemsets from other party, these itemsets are encrypted again and sent to a next party. This process is repeated until all itemset are encrypted by all parties.
3. A site, which is not site *D*, determines the union of all locally frequent itemsets.
4. CDK is calculated in order to decrypt itemsets found in step 3.
5. Using the itemsets from the previous step, globally frequent itemsets are found and sent to all parties. The algorithm terminates when there are no globally frequent itemsets.

Notice that the CDK decreases the number of expensive decryption operations required to obtain the result since every itemset is decrypted only once. More detailed information about the CDK can be found in [5].

3 SSAPP Algorithm

In contrast to the CDKSU algorithm, the SSAPP algorithm mines frequent sequences using a SS-Tree. Unlike an itemset, a sequence is an ordered list of items. Moreover, we assume that the same items may not appear multiple times at different positions in the sequence. Some key definitions are given below.

Definition 1
Given a distinct set of items $E = e_1, e_2, \ldots, e_n$, a sequence is defined as $\langle a_1 a_2 \ldots a_m \rangle$ where $a_i \in E$ ($1 \leq i \leq m$) and for any two items a_i and a_j ($i \neq j$) we have $a_i \neq a_j$.

A k-sequence is a sequence with length k, where the length of a sequence is the number of items in it.

Definition 2

A sequence $s_1 =< b_1 b_2 \ldots b_m >$ *is a subsequence of (i.e., is contained in) a sequence* $s_2 = \langle a_1 a_2 \ldots a_n \rangle$ *(*$n \geq m$*), denoted as* $s_1 \subseteq s_2$*, if there exist integers* $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ *such that* $b_1 = a_{i_1}$, $b_2 = a_{i_2}$,*. . .,* $b_m = a_{i_m}$. *On the other hand, the sequence* s_2 *is a supersequence of* s_1.

Definition 3

A i-prefix of a k-sequence s ($i \leq k$*) is a sequence that contains the first i items of s. Similarly, i-suffix of a k-sequence s (*$i \leq k$*) is a sequence that contains the last i items of s.*

Let s be the sequence ⟨U K A F⟩. Then, a 2-prefix of s is the sequence ⟨U K⟩ and a 2-suffix of s is ⟨A F⟩ . In particular, 4-prefix and 4-suffix are ⟨U K A F⟩.

Definition 4

The relative support supp of a sequence s is calculated as:

$$supp(s) = \frac{number\ of\ sequences\ containing\ s}{total\ number\ of\ sequences} * 100\%$$

Similarly, the number of sequences that contain s is called the absolute support (asupp) of s.

For instance, let a sequence database contain two sequences ⟨Z C A D⟩ and ⟨D A Z G⟩. The relative support of ⟨Z A D⟩ is 50% since only ⟨Z C A D⟩ matches ⟨Z A D⟩.

Definition 5

A sequence s is called frequent if its support supp (asupp) is no less than a threshold minSupp given by a user, i.e.:

$supp(s) \geqslant minSupp$ *(or* $asupp \geqslant minSupp$*)*

Definition 6

A sequence s is called locally frequent at site S_i *if its support is no less than a threshold minSupp with respect to* DB_i.

Definition 7

A sequence s is called globally frequent if its support is no less than a threshold minSupp with respect to DB, where $DB = \bigcup_{i=1}^{n} DB_i$ *and n is the number of parties.*

Definition 8

The placeholder x for a 2-sequence is a unique identifier which is defined as $x=\langle a_1 a_2 \rangle$ *(i.e.,* $\langle a_1 a_2 \rangle$ *can be replaced by x and vice versa).*

For instance, using the placeholder $x=\langle a_1 a_2 \rangle$ the sequence $s=\langle a_1 a_2 a_3 \rangle$ can be written as $s=\langle x a_3 \rangle$.

It should be emphasized that sequential patterns are essentially different from frequent itemsets. For instance, the number of result patters may be different for the same threshold and database.

3.1 SS-Tree

The SSAPP algorithm extracts frequent sequences from a prefix tree called SS-Tree. The SS-Tree is a compact data structure that is built in a single pass over a sequence database and consists of nodes that correspond to items. Furthermore, each node holds information about its frequency in a path in the tree and has a list of pointers (possible empty) to its children.

In order to indicate that a node matches an item, the item identifier is assigned to the node. Notice that an identifier of an item is unique (see Definition 1), however it can be used in many different nodes, as well as in sequences.

To enable a quick access to a list of nodes with a given identifier, an auxiliary data structure called a header table *ht* is maintained. Each element in this table has the identifier of an item *e*, counter that keeps track of the total number of occurrences of *e* in a SS-Tree and a list of pointers to nodes that correspond to *e*.

Additional definitions are as follows. A root is a node with no parent, and leaf is a node with no children. A subtree of a tree *t* is a tree consisting of a node (treated as a root) and all of its descendants in *t*. A path in a tree (or subtree) is a sequence of nodes from a root to a given node.

3.2 SS-Tree Construction

The construction algorithm of the initial SS-Tree is as follows. During a database scan, one sequence is read at a time and its items are mapped to a path in the initial tree. A sequence is inserted starting from a root node (denoted as *null*) of the initial tree. If the prefix of a sequence overlaps an existing path in a tree, the counter of each node in that path is incremented by 1; for unmatched items (i.e., items forming a suffix of the sequence) new nodes are created and their counters are initialized to 1. In particular, if a sequence fully overlaps an existing path no new nodes are added and counters of all nodes in that path are incremented by 1; on the contrary (i.e., if there is no path in a tree) for all items in the sequence new nodes are created and their counters are set to 1. In a similar way the header table is updated when sequences are inserted into the initial tree. The construction process continues until all sequences are added to the initial tree.

Figure 1 shows the SS-Tree after inserting all sequences from Table 1. Pointers stored in our header table are represented by dotted lines.

For more information about a similar tree, please see [8].

Table 1. Sequence database

No.	input sequence
1	O→ G→ F→ R→ B
2	B→ H→ J→ K
3	H→ G→ J→ K
4	B→ G
5	O→ G→ F → J
6	B→ H→ J

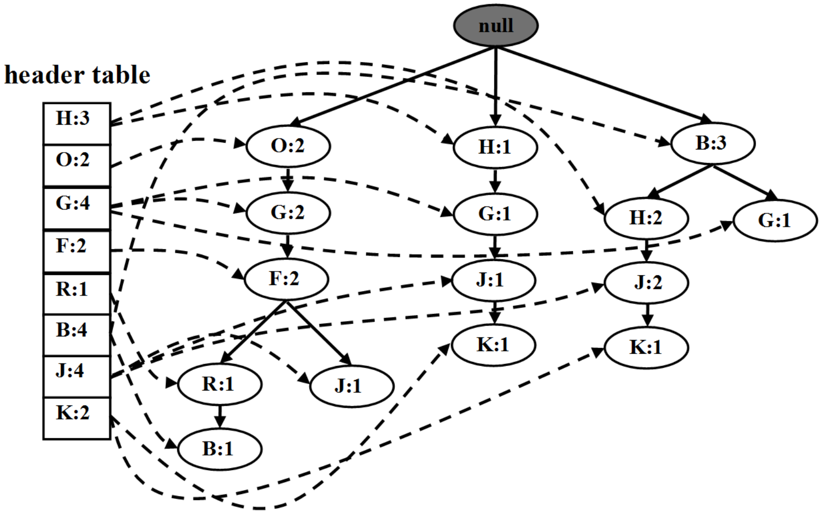


Fig. 1. Initial SS-Tree

3.3 SSAPP Algorithm

The process of mining frequent sequences can be seen as first mining frequent 1-sequences and then progressively growing them. In this view, in each iteration the length of frequent sequences is increased by 1. For instance, frequent 1-sequences are obtained in the first iteration, frequent 2-sequences in the second iteration and so on. Moreover, an iteration consists of a sequence of steps, i.e, rebuilding of current SS-Tree, extraction of frequent sequences and result forwarding.

3.4 Mining Frequent Sequences

In order to simplify our further discussion, we assume, without loss of generality, that each site (data owner) has the same sequence database. This implies that locally frequent sequences are also frequent globally.

Another assumption is that a site that broadcasts results to all the sides adds one additional item to every globally frequent sequence. This auxiliary item is called a placeholder (see Definition 8) and it can be seen as a unique identifier of a globally frequent sequence. Since a placeholder substitutes a 2-sequence, it is added only when the length of a globally frequent sequence is 2. In consequence, every globally frequent sequence received by a site consists of either a single item (first iteration) or two items and a placeholder (next iterations).

Let us examine two examples. After receiving all globally frequent sequences, each site uses placeholders to rebuild a SS-Tree from the previous iteration. Assuming that we have a sequence database as shown in Table 1 and the minimum support threshold is set to 2, the following globally frequent sequences are obtained in the second iteration of the SSAPP algorithm: {⟨G F⟩, X1}, {⟨G J⟩, X2}, {⟨H J⟩, X3}, {⟨H K⟩, X4}, {⟨O F⟩, X5},

{⟨O G⟩, X6}, {⟨J K⟩, X7}, {⟨B H⟩, X8}, {⟨B J⟩, X9}. Please notice that placeholders are denoted with a capital letter X followed by a number. Furthermore, Figures 2-3 present three subtrees of the initial SS-Tree. As we can see in these figures, G is highlighted in gray which means that only globally frequent sequences having G as the first item will be involved in our examples, i.e., ⟨G F⟩, ⟨G J⟩.

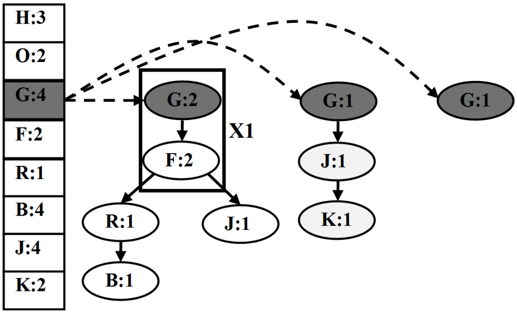


Fig. 2. Example for placeholder X1

Example 1. According to Definition 8, X1 is a a substitute (i.e., placeholder) for ⟨G F⟩ and we can write X1=⟨G F⟩. Since ⟨G F⟩ can be replaced by a single item X1, all sequences starting with ⟨G F⟩ have to be found in the SS-Tree. In Figure 2, there is only one subtree that matches ⟨G F⟩; matched sequence in the subtree is denoted using a rectangle. Moreover, subtrees ⟨G J K⟩ and ⟨G⟩ are not involved in our analyze since they do not have items associated with F. As a result, the placeholder X1 has two child nodes R and J.

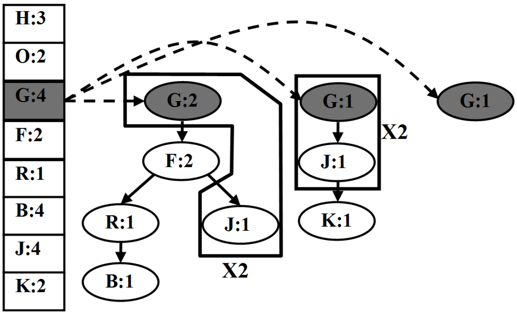


Fig. 3. Example for placeholder X2

Example 2. In a similar way as before, we can write X2=⟨G J⟩. In this case there are two subtrees (see Figure 3) that match ⟨G J⟩. ⟨G F J⟩ is a valid sequence since, according

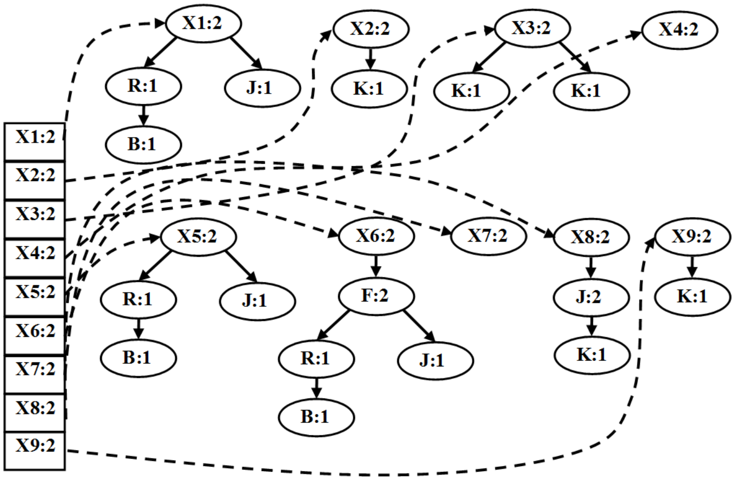


Fig. 4. SS-Tree built in the second iteration

to Definition 2, ⟨G F J⟩ is a supersequence of ⟨G J⟩. The placeholder X2 has only K as a child node.

Transformations of the other placeholders are analogous to those presented in Figures 2-3.

Figure 4 shows the SS-Tree after the rebuilding process. Please note that not all of the nodes have counterparts in the new header table. This is because, starting from second iteration, the SSAPP algorithm will only keep track of occurrences of placeholders, which can be thought of as tracking nodes whose children should be considered in next iterations.

In the next step, the SSAPP algorithm extracts all frequent 2-sequences that start with a placeholder, i.e., ⟨X3 K⟩, ⟨X6 F⟩, ⟨X8 J⟩. These frequent sequences are sent to other site.

In the third iteration, after rebuilding the SS-Tree, the SSAPP algorithm terminates since no frequent sequence can be generated.

3.5 Pseudocode

The pseudocode of the SSAPP algorithm is given below. Notice that the SSAPP algorithm is performed by each data owner.

Details about communication between sites are omitted since they are already described in Subsection 2.1. An important difference is that one site adds a placeholder to every globally frequent sequence sent to all sites.

4 Experiments

In the experiments, we compared two algorithms SSAPP and AprioriAll [9]. The AprioriAll algorithm is equivalent to the SSAPP algorithm, since it generates the same result set for a given *minsupp* threshold. Both algorithms are embedded into CDKSU,

Algorithm 1. SSAPP algorithm

```
1) if initialization then
2)     Construct an initial SS-Tree
3) else
4)     if set of globally frequent sequences is empty then
5)         Terminate the algorithm
6)     else
5)         Output globally frequent sequences
6)         Based on placeholders, rebuild SS-Tree
7) Using the SSAPP algorithm, extract locally frequent sequences from the SS-Tree
8) Send locally frequent sequences to other site
```

but a new strategy is only used in the SSAPP algorithm. A commutative encryption scheme used in the experiments is based on Elliptic Curve Pohlig-Hellman cipher with prime239v1 parameters [10].

The experiments were carried out on 4, 5, 6 and 7 nodes (computers) connected by a local area network. Each of the nodes had a sequence database consisting of 10k sequences with 629 distinct items. The average length of sequences was 14.2 and the average deviation was 0.5. Table 2 shows a summary of results (minimum support threshold, total number of patterns, average length and average deviation of patterns). All computers were equipped with the Intel Xeon W3550 @ 3.07GHz processor running on Microsoft Win 7 Prof SP1, 4GB of RAM and Seagate Barracuda 7200.12 500GB hard drive.

Table 2. Summary of results

minsupp	patterns	avg. length	avg. deviation
1.5%	6171	4.15	1.4
1.0%	39360	5.5	1.58
0.5%	814440	7.52	1.7

The runtime of the algorithms for different values of the minimum support is shown in Figure 5. The suffix *en* means that encryption is on, and *no* that there is no encryption involved.

The results given in Figure 5 (left side) show that the total runtime (expressed in a logarithmic scale on the vertical-axis) increases when decreasing a minimum support threshold from 1.5% to 0.5%. In all cases, the SSAPP algorithm performs much better than AprioriAll.

Figure 5 (right side) shows the runtime of the algorithms (without encryption, decryption and communication costs). As seen from the results shown in this figure, the SSAPP algorithm greatly outperforms the AprioriAll algorithm.

The runtime of SSAPP and AprioriAll as the support threshold decreases from 1.5% to 1.0% is shown in Figure 6. The number of nodes involved in the experiment is denoted in the suffix of the algorithm's name. We can observe that the runtime increase is nearly constant per node, however the encryption and decryption overheads are lower

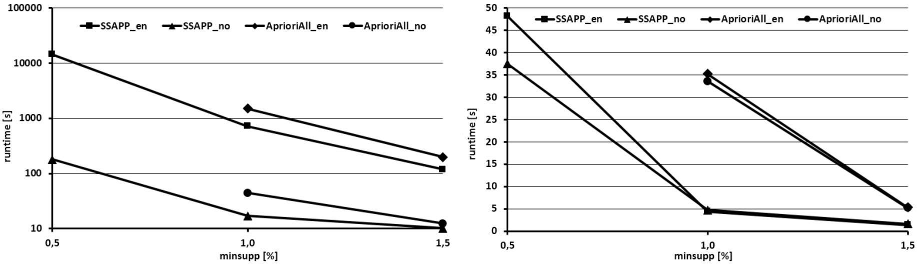


Fig. 5. Scalability against threshold (4 computers). Total runtime (left side) and runtime of algorithm (right side).

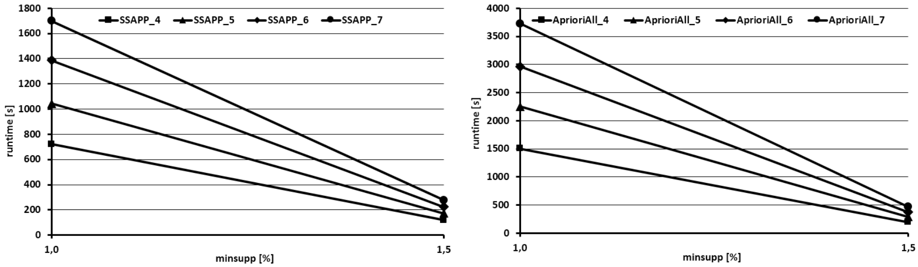


Fig. 6. Total runtime vs. number of nodes

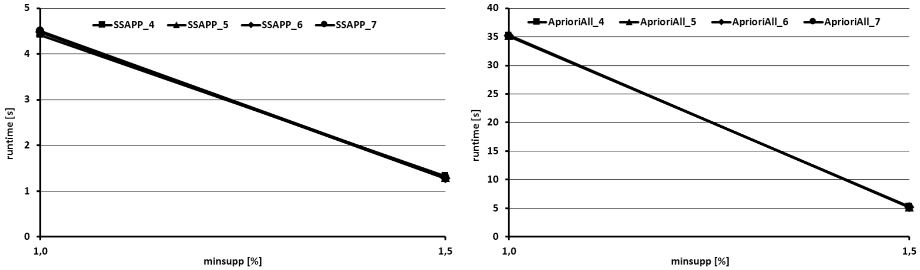


Fig. 7. Runtime of algorithms (without encryption, decryption and communication) vs. number of nodes

in the SSAPP algorithm. Notice that in the case of the AprioriAll algorithm, when the support threshold is lower than 1.0% and encryption is on, patterns cannot be generated within reasonable time.

As expected, see Figure 7, the runtime (without encryption, decryption and communication costs) of both algorithms is constant for a given support threshold and does not depend on the number of nodes.

From the experiments, one can see that the AprioriAll algorithm is slower than the SSAPP algorithm.

5 Conclusion

The paper presents the SSAPP algorithm for sequential pattern mining in a distributed environment with privacy-preserving. This algorithm arranges sequences in a tree called the SS-tree.

The SSAPP algorithm allows to reduce the number of data blocks that are encrypted and decrypted using a commutative cipher. The experiments show that reduction of operations can improve performance significantly.

In the future it is planned to investigate the proposed mechanism in a more comprehensive way. Moreover, the SSAPP algorithm will be integrated into the Trajectory Data Warehouse [11].

References

1. Goldreich, O., Micali, S., Wigderson, A.: How to Play any Mental Game or A Completeness Theorem for Protocols with Honest Majority. In: STOC, pp. 218–229 (1987)
2. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: KDD, pp. 639–644 (2002)
3. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, M.Y.: Tools for Privacy Preserving Data Mining. SIGKDD Explorations, 28–34 (2002)
4. Kantarcioglu, M., Clifton, C.: Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. In: DMKD, pp. 1026–1037 (2002)
5. Gorawski, M., Siedlecki, Z.: Optimization of Privacy Preserving Mechanisms in Homogeneous Collaborative Association Rules Mining. In: ARES, pp. 347–352 (2011)
6. Gorawski, M., Jureczek, P.: Optimization of privacy preserving mechanisms in mining continuous patterns. In: Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J. (eds.) New Results in Dependability & Comput. Syst. AISC, vol. 224, pp. 183–194. Springer, Heidelberg (2013)
7. Gorawski, M., Jureczek, P.: Extensions for Continuous Pattern Mining. In: Yin, H., Wang, W., Rayward-Smith, V. (eds.) IDEAL 2011. LNCS, vol. 6936, pp. 194–203. Springer, Heidelberg (2011)
8. Gorawski, M., Jureczek, P.: Continuous Pattern Mining Using the FCPGrowth Algorithm in Trajectory Data Warehouses. In: Graña Romay, M., Corchado, E., Garcia Sebastian, M.T. (eds.) HAIS 2010, Part I. LNCS, vol. 6076, pp. 187–195. Springer, Heidelberg (2010)
9. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: ICDE, pp. 3–14 (1995)
10. Qu, M.: Standards for efficient cryptography sec 2: Recommended elliptic curve domain parameters (2010)
11. Gorawski, M., Jureczek, P.: Regions of Interest in Trajectory Data Warehouse. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010, Part I. LNCS, vol. 5990, pp. 74–81. Springer, Heidelberg (2010)

User Identity Unification in e-Commerce

Marcin Gorawski[1,2], Aleksander Chrószcz[2], and Anna Gorawska[2]

[1] Wroclaw University of Technology,
Institute of Computer Science,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
Marcin.Gorawski@pwr.wroc.pl

[2] Silesian University of Technology,
Institute of Computer Science,
Akademicka 16, 44-100 Gliwice Poland
{Marcin.Gorawski,Aleksander.Chroszcz,Anna.Gorawska}@polsl.pl

Abstract. Data mining applied to social media is gaining popularity. It is worth noticing that most e-commerce services also cause the formation of small communities not only services oriented toward socializing people. The results of their analysis are easier to implement. Besides, we can expect a better perception of the business by its own users, therefore the analysis of their behavior is justified. In the paper we introduce an algorithm which identifies particular customers among not logged or not registered users of a given e-commerce service. The identification of a customer is based on data that was given so as to accomplish selling procedure. Customers rarely use exactly the same identification data each time. In consequence, it is possible to check if customers create a group of unrelated individuals or if there are symptoms of social behavior.

1 Introduction

When we posses a complete database of selling transactions then we can apply a group of well known algorithms[19] such as the analysis of shopping basket contents or the analysis of the most frequently sold items. Problems emerge when the volume of customers and transactions is small. Then, there is a huge risk that the current data excludes some groups of people which contain potential new customers. The weakness of the above methods manifests itself also in the case of frequently changing ranges of goods. This happens on the pharmaceuticals & cosmetics market which rapidly implements scientific discoveries. As a result, historical sales analysis has little usefulness when new products are considered.

In this area, multi dimensional data mining of social networks[22], news portals[2] and auction portals has become a promising solution. Thanks to a huge volume of users registered in such services, those systems are considered to be an unbiased opinion centers. There are projects extracting hot topics from Twitter messages by means of text analysis[22]. Other projects are designed to trace knowledge propagation[1] in Internet communities. As a result, we can gain access to opinions about new products and advertising campaigns.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_16, © Springer International Publishing Switzerland 2014

Social analysis are valuable also for smaller e-commerce systems. The paper addresses the problem of customer identification in selling transactions collected from some e-shopping system. The aim is to recreate customer list and assign transactions to them. The incomplete attribute set constitutes the main challenge. Customers which are not registered leave a lot of attributes empty. It was popular that they fill in e-mails or telephone numbers exclusively. Besides it is possible that customers use data which does not identify them. For instance the shipping address may not be the customer's own address. In order to solve the above task, we have designed a clustering algorithm which groups together transactions committed by the same customer. In contrast to a typical clustering task, this solution is expected to generate thousands of clusters. In consequence, there is required the algorithm which automatically estimates the cluster number. Moreover, the distance function applied to compare the similarity between clusters should work even some attribute values are missing. Those specific requirements inspired us to create a new algorithm. Among potential strategies [13,16,23,4] we have chosen agglomerative one because this algorithm can be adapted to work with our custom function which measures distance between clusters.

This problem appeared in a real system. We have an opportunity to carry out data cleaning and data integration for a medium sized e-commerce service existing for a few years. In order to attract reluctant customers, this e-commerce systems do not oblige all customers to create their account. Particularly infrequent or sporadic users do not create accounts. On the other hand, knowing their profiles would enable that company to acquire them as a regular customers.

The paper is organized as follows: Section 2 describes the data structure of processed data. In Section 3 our clustering algorithm is defined. Next, Section 4 contains conducted experiments. Than some aspects of confidential data leakage are discussed in order to show potential risks connected with a customer behavior analysis. Finally, in Section 6 we conclude our results.

2 Data Structure

Let us assume that there are two tables: *User* and *Transaction* depicted in fig. 1. The *Transaction* table defines which user in a given day has bought goods. Each transaction is accompanied by attributes describing billing and shipping address details like: post code, country, state and address. In order to make the model simple, the list of products in transactions is omitted. Because the customer's address, e-mail and identification attributes may change in time for a given customer, the valid values are stored separately in the *transaction* table when a transaction is submitted.

In order to check if two transactions were concluded by the same customer it is necessary to compare the transaction attributes. As we can see, the data scheme allows us to store information about transactions on low granularity level. On the other hand, only a narrow group of attributes are mandatory. As a result, a considerable number of null values exist in the database. Especially, billing

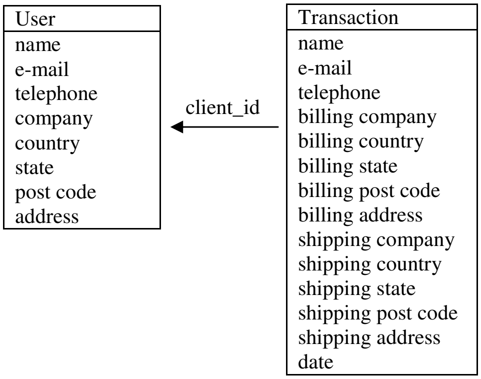


Fig. 1. Schema of data source

addresses details were rarely filled in the datasets on which we have worked. Bellow we describe other data imperfections which were present in the processed data.

It may seam that the *name* and *company* name will be helpful in the identification of a particular customer. In reality, customers are usually not consistent and use plenty of abbreviations when the name is long.

In contrast to previous attributes, e-mail values were always verified in the system. Despite this fact, particular customers cannot be identified by means of this attribute only for two reasons: customers usually have a few e-mails; besides, a group of transactions are concluded on behalf of somebody. This occurs when office work is delegated to subordinates. Such information may be strategic, therefore, we will also consider this aspect in the paper.

Another noteworthy thing is that a telephone number combines a few pieces of information. A complete telephone number is useful when it is a personal telephone number. On the other hand, we can analyze only the prefix. In consequence, we can identify with high probability people coming from the same company.

The limitations of address usage are similar to those diagnosed for *company* name and customer *name* attributes. Therefore, it is important to unify those values before comparison. It was common in the analyzed data that a customer sometimes inserts unnecessarily a room or suite number, dots, comas and empty spaces. In order to make the address unification process more resistant to possible textual variations, we have prepared a dictionary of popular abbreviations. In consequence, we can arrive at an address with all the possible abbreviations applied regardless of which address version the algorithm started with.

3 Algorithm

Our clustering algorithm is a combination of agglomerative clustering [21,3,25] with a distance function based on the Naive Bayesian classifier[20]. The other solutions like k-means and canopy [23] clustering are better suited for problems with fewer clusters.

Our clustering process is divided into two phases. During the first phase, the Naive Bayesian classifier is taught. It checks if two transactions were concluded by the same customer. The answer is expressed as:

$$similarity = pg/(pg + npg) \tag{1}$$

where:

- pg - the probability that two clusters of transactions were concluded by the same customer,
- png - the probability that two clusters of transactions were concluded by different customers.

When *similarity* is grater than 0.5 it means that both transactions belong to the same customer. Values pg and png approximate the probabilities defined above. Therefore it is necessary to use the *similarity* value which normalizes pg according to $pg + npg$. The algorithm of pg and png estimation will be explained in subsection 3.2.

The next phase implements the agglomerative algorithm which combines the most similar transactions until there are no clusters with similarity greater than 0.5.

3.1 Agglomerative Clustering

Agglomerative clustering [24] is a greedy algorithm which implements a bottom-up strategy. Each observation starts in its own cluster, then according to a distance function the closest pairs of clusters are merged into a single larger cluster. The process repeats until a single cluster is created. The customizations of this algorithm are used in a wide range of applications and fields[21,4,18,25] including data mining, compression and image processing.

In order to optimize the basic algorithm version, we have introduced a heap which preserves information on the similarity of transactions [24]. The resulting algorithm is shown in alg. 1. Function $findClosestMatch(A)$ returns cluster B whose pair (A, B) maximizes *similarity* between all transactions with the restriction $A \neq B$.

At the beginning, the heap is initialized with pairs of the most similar clusters. Each cluster has the most similar one assigned to it. When new cluster C is created from clusters A and B then pairs containing either A or B are no longer valid in the heap. Those entries can be immediately removed. The introduced algorithm removes those elements lazily. Each time another element from the heap is processed, it is checked if any cluster from the pair has already been incorporated into a bigger cluster and removed from the *clusters* list. When this is true, the further analysis of this element is omitted. The lazy strategy simplifies the algorithm and makes it faster.

Algorithm 1 *Heap-based agglomerative clustering.*

```
clusters = list of transactions in their own clusters
MinHeap heap = new MinHeap();
for(A in clusters) {
  Cluster B = findClosestMatch(A);
  heap.add(new Pair(A,B));
}

while(!heap.isEmpty()) {
  Pair pair = heap.poll();
  if(pair != null) {
    Cluster tmpClL = null;
    Cluster tmpClR = null;
    if((tmpClL = clusters.get(pair.cluster1Id))==null) {
      //tmpClL was already clustered with somebody
    } else if((tmpClR = clusters.get(pair.cluster2Id))==null
        ) {
      Pair newPair = findClosestMatch(tmpClL, clusters);
      if(newPair != null)
        heap.add(newPair);
    } else {
      if(pair.distance > 0.5) {
        clusters.remove(tmpClL.id);
        clusters.remove(tmpClR.id);
        Cluster C = new Cluster(tmpClL, tmpClR);
        clusters.put(C.id, C);
        Pair newPair = findClosestMatch(C, clusters);
        if(newPair != null)
          heap.add(newPair);
      }
    }
  }
}
```

3.2 Distance Function

We have decided to measure the similarity between transactions by means of the Naive Bayesian classifier because it allows us to compare attributes with missing values. Let us assume that we have two transactions to be classified. At the beginning, the corresponding attributes of those transactions are compared and the result is stored in new record R. The comparison of two corresponding attributes may yield one of the three situations: equal, not equal and unknown. The last outcome occurs when at least one transaction has no defined value for the compared attributes. According to values stored in R, the Naive Bayesian classifier calculates the probability of concluding transactions by the same customer.

This classifier is based on the bellow probability model:

$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C}{p(F_1, ..., F_n)} \tag{2}$$

The variable C has two states which indicate whether two transactions were concluded by the same customer or not. Variables $F_1, ..., F_n$ represent similarity of transaction attributes. This model leads to the following Naive Bayesian classifier rule:

$$\begin{array}{l} classify(f_1, ..., f_n) = \\ \underset{c}{\operatorname{argmax}}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c) \end{array} \tag{3}$$

The conditional probabilities $p(F_i = f_i | C = c)$ are measured in the learning phase. Because this task is not complicated, we don't explain those calculations in detail. Let us focus on probability $p(C)$. We want to calculate the probability of two random transactions being concluded by the same customer. When the probability of a transaction belonging to customer A appears in the database with probability p_A, we can conclude that two randomly selected transactions belong to the same customer with probability p_A^2. This feature explains why we cannot measure $p(C)$ on the basis of a learning dataset. This dataset is a fraction of a complete database. In consequence, it contains a different number of customers in comparison with the complete dataset. We have designed the following algorithm to measure the average probability $p(C = true)$ that two transactions being concluded by the same customer. Let us assume that the size of the dataset to be clustered equals $sizec$ and the average number of transactions per customer in the learning dataset equals avt. Then the probability $p(C = true)$ is approximated as the multiplication of the average customer number $sizec/avt$ by the probability of the pairs of their transactions being randomly chosen $(avt/sizec)^2$. Summing up the approximated probability $p(C = true)$ equals:

$$p(C = true) = \frac{avt}{sizec} \tag{4}$$

In contrast to the classifier definition, the distance function operates on clusters not on single transactions. In order to adapt the Naive Bayesian classifier, we have decided to measure: a) the maximum probability *pg* of two clusters A and B contain transactions concluded by the same customer; b) the maximum probability *npg* of two clusters contain transactions concluded by different customers. Then those values are used to evaluate the *similarity* variable.

Let us consider the dataset showed in tab. 2. Columns Tr_Id and Cl_Id represent the transaction key and the cluster key respectively. Cluster 1 and cluster 2 do not have one unique value for attributes $e - mail$ and $phone$. This property prevents a straightforward application of the Naive Bayesian classifier. Having analyzed this dataset, we cannot identify pair (A, B) where A represents a transaction from cluster 1 and B stands for a transaction from cluster 2 which have equal e-mail and phone values. On the other hand, when transactions 4 and 5 are clustered together, it means each value of a given attribute identifies the cluster. As a result, the pair (*bob@domain.com*, 123456797) identifies cluster 2. This observation leads to the solution, clusters A and B are equal on the level of a given attribute if there is at least one value which exists in both clusters. On the other hand, the dissimilarity distance between clusters equals maximum *npg* between transactions belonging to different clusters.

Table 1. Example transactions

Tr_Id	Cl_Id	E-mail	Phone	Name
1	1	clara@domain.com	123456797	Bob Hopkins
2	1	clara@domain.com	123456798	Bob Hopkins
3	1	bob@domain.com	123456798	Bob Hopkins
4	2	adrian@domain.com	123456797	Adrian Jones
5	2	bob@domain.com	987654321	Mary Jones

Table 2. Clusterization comparison

Attribute	Test1	Test2
Size	10221	4501
Basic	1526	1732
Cluster	2944	1626

4 Experiments

At the beginning we have created a simple method which identifies particular customers among selling transactions according to client names and e-mails. This method quality was low because it is not resistant to any variations of textual data. On the other hand, it was sufficient to identify following categories: a) customers with one or a few e-mail addresses; b) e-mail addresses assigned to the group of customers; and c)transactions where e-mail addresses and customer names create complicated relations. We have diagnosed that this method creates too many partitions for categories a) and b). However the most problematic was category c) which creates partitions containing many customers. In consequence, they were useless to any further analysis. Then we have applied the introduced clustering algorithm.

The test was conducted on the computer with 4GB RAM and Intel Core 2 Duo P8600 2.4GHz. Our dataset was a real data set of around 35 000 transactions coming from e-commerce system. We have created two testing data sets. $Test1$ consists of all transactions belonging to the category c). $Test2$ contains 4 501 transactions belonging to categories a) and b). Because partitions belonging to data category a) and b) identify particular customers almost correctly, we have decided to use 4 000 transactions coming from those partitions as a learning dataset without additional processing. Next we have processed datasets $Test1$ and $Test2$. Then the resulting clusters were verified manually by the client if they meet the expectations. Table 2 concludes our results. The $Size$ attribute is the dataset size. Basic and Cluster refers to the number of identified particular customers by means of the basic partitioning algorithm and the agglomerative version respectively.

We can see that our solution achieve both aims. Our cluster algorithm almost doubles the amount of particular customers for category c). It also creates less clusters for datasets a) and b). It is noteworthy that the testing data was extracted automatically which allows us to automatize the whole process. Additionally, we have measured the calculation time of clustarization. The datasets *Test*1 and *Test*2 were processed during 14 minutes and 3 minutes respectively. The fist dataset was processed noticeably longer because it contains more similar transactions. Nevertheless, the time complexity of our algorithm is adequate to the data available in medium sized e-commerce database.

5 Confidential Data Leakage

Transactions X & Y concluded by 2 different customers may, however, be clustered together. This data weakness introduces an error into the further analysis of customer behavior. The less obvious and more serious consequence of this inaccuracy is the risk of confidential data leakage.

Let us assume that a marketing campaign aimed at narrow groups of customers was created. Next, leaflets are sent out to customers stored in the reconstructed customers database. When a leaflet is sent to the address specified in transaction X as well as at the one specified in transaction Y, then those two customers get some confidential data about each other. For instance, one can get unauthorized access to the information what range of products is popular with the other party. To be on the safe side, sending leaflets only to one address per customer is recommended.

Unfortunately the above suggestion will still permit the following scenario. Let us assume that unregistered customer A uses a given e-commerce service on a regular basis and we want to invite this user to join a loyalty program. Unfortunately, transactions from customers A and B were merged during clusterization by mistake. If this invitation were sent to customer B, then he would get a loyalty program suited for customers regularly purchasing product X. Because the relation between customers A and B must be close, customer B may guess that this loyalty program should have been sent to A. In consequence, customer B gets some confidential data while customer A is not aware of this fact.

6 Conclusion

There are powerful methods for analyzing the behavior of users in social networks. These methods help drawing conclusions about the attitude of users to products. However, many customer databases are sparsely populated. Even worse, the data in such databases may be ambiguous because the same person uses different proxies for his/her activities. Moreover, the name of a single customer may have been recorded in many slightly different ways. Hence, we need a method that assigns to each customer the data really associated to him/her, using a unique terminology.

The ability to work on real data shows that transaction log from e-commerce has hidden information about customers relationship. Those relations appear when we analyze e-mails, addresses and phone prefixes because institutions usually have constant elements. On the other hand, people tend to use their own data even when they make transactions for a company.

In the paper, we have proposed the new algorithm which is a composition of agglomerative clustering and the Naive Bayesian classifier. Thanks to it we can uncover unique customers from the concluded transactions. In the future work, we want to optimize this algorithm calculation complexity and adapt it to data warehouse systems [14,11,15,17,5,10,9,12,6,7]. Another interesting aspect of further research is as follows. Having known the relations between customers, we can explain the propagation of customer actions using the terminology of stream processing data. Our experiments in the area of stream database optimizations [8] show that the partitioning of data processing plan reduces substantially information latency. When customers are assumed as data stream operators then we can use similar strategy to accelerate information propagation in such a graph of customer relationship. In consequence, the communication between customer and e-commerce may gain from the better understating of users' interaction habits.

References

1. Asur, S., Huberman, B.A., Szabó, G., Wang, C.: Trends in social media: Persistence and decay. CoRR, abs/1102.1402 (2011)
2. Awadallah, R., Ramanath, M., Weikum, G.: Opinionetit: understanding the opinions-people network for politically controversial topics. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 2481–2484. ACM, New York (2011)
3. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proceedings of the Sixth ACM SIGKDD on Knowledge Discovery and Data Mining, KDD 2000, pp. 407–416. ACM, New York (2000)
4. Berkhin, P.: Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA (2002)
5. Gorawski, M.: Extended cascaded star schema and ECOLAP operations for spatial data warehouse. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 251–259. Springer, Heidelberg (2009)
6. Gorawski, M., Bańkowski, S., Gorawski, M.: Selection of structures with grid optimization, in multiagent data warehouse. In: Fyfe, C., Tino, P., Charles, D., Garcia-Osorio, C., Yin, H. (eds.) IDEAL 2010. LNCS, vol. 6283, pp. 292–299. Springer, Heidelberg (2010)
7. Gorawski, M., Chrószcz, A.: Streamapas: Query language and data model. In: CISIS, pp. 75–82 (2009)
8. Gorawski, M., Chrószcz, A.: Optimization of operator partitions in stream data warehouse. In: Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP, DOLAP 2011, pp. 61–66. ACM, New York (2011)
9. Gorawski, M., Chrószcz, A.: Synchronization modeling in stream processing. In: Morzy, T., Härder, T., Wrembel, R. (eds.) ADB15. AISC, vol. 186, pp. 91–102. Springer, Heidelberg (2013)

10. Gorawski, M., Gorawski, M.: Balanced spatio-temporal data warehouse with r-mvb, stcat and bitmap indexes. In: PARELEC, pp. 43–48 (2006)
11. Gorawski, M., Gorawski, M.: Modified R-MVB tree and BTV algorithm used in a distributed spatio-temporal data warehouse. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Wasniewski, J. (eds.) PPAM 2007. LNCS, vol. 4967, pp. 199–208. Springer, Heidelberg (2008)
12. Gorawski, M., Jureczek, P.: Continuous pattern mining using the FCPGrowth algorithm in trajectory data warehouses. In: Graña Romay, M., Corchado, E., Garcia Sebastian, M.T. (eds.) HAIS 2010, Part I. LNCS, vol. 6076, pp. 187–195. Springer, Heidelberg (2010)
13. Gorawski, M., Malczok, R.: AEC algorithm: A heuristic approach to calculating density-based clustering *Eps* parameter. In: Yakhno, T., Neuhold, E. (eds.) ADVIS 2006. LNCS, vol. 4243, pp. 90–99. Springer, Heidelberg (2006)
14. Gorawski, M., Malczok, R.: AEC algorithm: A heuristic approach to calculating density-based clustering *Eps* parameter. In: Yakhno, T., Neuhold, E. (eds.) ADVIS 2006. LNCS, vol. 4243, pp. 90–99. Springer, Heidelberg (2006)
15. Gorawski, M., Malczok, R.: Materialized ar-tree in distributed spatial data warehouse. Intell. Data Anal. 10(4), 361–377 (2006)
16. Gorawski, M., Malczok, R.: Towards automatic *eps* calculation in density-based clustering. In: Manolopoulos, Y., Pokorný, J., Sellis, T. (eds.) ADBIS 2006. LNCS, vol. 4152, pp. 313–328. Springer, Heidelberg (2006)
17. Gorawski, M., Marks, P.: Checkpoint-based resumption in data warehouses. In: Sacha, K. (ed.) SET. IFIP, vol. 227, pp. 313–323. Springer, Heidelberg (2006)
18. Guha, S., Rastogi, R., Shim, K.: Cure: an efficient clustering algorithm for large databases. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD 1998, pp. 73–84. ACM, New York (1998)
19. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)
20. Heller, K.A., Ghahramani, Z.: Bayesian hierarchical clustering
21. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31(3), 264–323 (1999)
22. Mathioudakis, M., Koudas, N.: Twittermonitor: trend detection over the twitter stream
23. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the Sixth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2000, pp. 169–178. ACM, New York (2000)
24. Olson, C.F.: Parallel algorithms for hierarchical clustering. Parallel Computing 21, 1313–1325 (1993)
25. Walter, B., Bala, K., Kulkarni, M., Pingali, K.: Fast agglomerative clustering for rendering. In: IEEE Symposium on Interactive Ray Tracing (RT), pp. 81–86 (August 2008)

Polarity Lexicon for the Polish Language: Design and Extension with Random Walk Algorithm

Konstanty Haniewicz[1], Monika Kaczmarek[1], Magdalena Adamczyk[3], and Wojciech Rutkowski[2]

[1] Department of Information Systems, Faculty of Informatics and Electronic Economy, Poznań University of Economics
{konstanty.haniewicz,monika.kaczmarek}@ue.poznan.pl
[2] Ciber Poland, Poznan, Poland
wojciech.rutkowski@ciber.com
[3] Department of Modern Languages, University of Zielona Góra
m.adamczyk@wh.uz.zgora.pl

Abstract. Sentiment analysis aims at an automatic assignment to a portion of text a value expressing an emotional attitude towards its content. Out of numerous efficient methods for investigating sentiment, the authors decided to opt for the lexicon-based approach. A necessary prerequisite for adopting it was the availability of specific lexical resources for the investigated language. While there are substantial readily accessible polarity resources for English, those for Polish are meagre and, to the best of our knowledge, none of them is able to fully support sentiment analysis. Accordingly, the main objective of the presented work is to plug this gap in academic research by creating in an automated manner, a polarity lexical resource for the Polish language. In this paper, we present the motivation for the study and the key mechanisms underlying the development of the polarity lexicon, elucidate the linguistic phenomena to be reckoned with in the process, as well as discuss the random walk algorithm used to extend the obtained polarity resources. Finally, the results of the conducted experiments and the newly compiled polarity lexicon are demonstrated.

Keywords: sentiment analysis, polarity lexical resources for Polish, Natural Language Processing, Random Walk Approach

1 Introduction and Motivation

The main task of sentiment analysis is to assign polarity to a text or its portion [1]. Sentiment analysis allows one to evaluate in an automated manner a huge wealth of information [2], which is why it can be of considerable benefit to users and organizations. For instance, it can provide data on whether actions taken to foster the recognisability of a brand, product or service yield the desired results. In addition, sentiment analysis is invaluable for quickly gauging the overall attitude of the Internet media towards a given topic.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_17, © Springer International Publishing Switzerland 2014

As mentioned above, the use of the lexicon-based approach to sentiment classification is preconditioned by the availability of relevant lexical resources, namely dictionaries of sentiment words providing for each item in a dictionary its sentiment score in a given domain. The lexical resources for Polish are pretty scarce and, to the best of our knowledge, none of them is capable of fully supporting sentiment analysis, as shown in the related work section below.

The overall aim of our work is to construct a large, free and general purpose polarity lexicon for the Polish language. The quality of the devised system is assumed to be strongly dependent on the quality of the developed resources. More specifically, if the created polarity semantic network (i.e., the sentiment lexicon) is not sufficiently extensive to satisfactorily cover the investigated language, the results of sentiment analysis may prove disappointing. Therefore, the first version of the sentiment lexicon based on the machine learning approach [3] was extended by means of applying the Markov random walk model making use of the available semantic lexicon, which resulted in producing a polarity estimate for a given word.

The presented research follows the pro-active research path based on the design oriented research paradigm [4] and the design science paradigm [5,6]. First, the concept building phase took place, which laid a sound theoretical foundation for our work. The next step was the approach building, which involved devising a reliable method of creating a polarity semantic network based on the previously formulated theoretical concepts. The developed method drew on a number of the already existing approaches to creating lexical resources. The subsequent stage was concerned with creating a semantic network with 140000 concepts, which was then extended by means of the random walk algorithm. Finally, experiments aimed at testing the quality of the developed artefact were conducted.

The paper is structured as follows. Initially, a brief overview of the research related to our work is presented.The next two sections are devoted to discussing the methods employed for developing the polarity semantic network for Polish and those adopted to extend the initial version of the compiled polarity lexicon. Subsequently, the conducted experiments are described in some detail. The paper concludes with final remarks and directions for future research.

2 Related Work

Sentiment analysis is defined as "the computational treatment of opinion, sentiment and subjectivity in text" [1] and consists of the following phases: part-of-speech tagging (division into language tokens); subjectivity detection (determining the statement as subjective or objective) and polarity detection (i.e., for subjective statements, the evaluation of their polarity) [2]. Since 2001 there has been a considerable growth in the interest in sentiment analysis [7], the reasons for which include the development of mature methods of analysing natural language, the availability of large volumes of test data on the Web and the increasing demand for intelligent applications [1]. Most studies on sentiment and opinion mining use English as the source of data. Yet, a growing number of research initiatives in the area are now drawing on the already existing resources for English to investigate these phenomena in other languages (e.g., Chinese [8]). While some efforts of this sort have also been made for Polish (e.g., [9]), they are altogether relatively scarce.

Essentially, there are two main approaches to sentiment classification, namely the supervised (machine) learning approach and the lexicon-based approach. The former makes use of the text classification framework and its most frequently applied type is the Support Vector Machines with n-gram features trained on a large set of texts with known polarities (usually positive or negative) (e.g., [10, 11]). While the machine learning techniques were shown to be suitable for sentiment analysis tasks (see [1]), further experiments (e.g., [12]) demonstrated that adopting the lexicon-based approach, making use of polarity lexicons, may significantly improve the results of the sentiment classification task.

Lexical resources cater for most of the linguistic information needed in Natural Language Processing and related areas [13]. The existing polarity lexical resources differ markedly in complexity. They may be lists of positive or negative words, as well as more complex semantic nets [14]. Moreover, as word sentiments are domain-specific [15], no general-purpose polarity lexicon is able to perform successfully for each domain [16]. That is why a number of domain-aware polarity lexicons have appeared, which substantially enhance the performance in sentiment classification tasks [16].

There are numerous polarity resources available for English, which include, among others, SentiWordNet [17], WordNet-Affec [18], ANEW [19]), as well as a number of general resources, such as WordNet [20] (for an in-depth overview of approaches to developing lexical resources see [1]). Creating a polarity lexicon manually is a labour-intensive and error-prone task, where a sufficiently extensive use of the lexicon cannot be guaranteed [16]. This explains why some research initiatives are designed to create polarity lexicons in an automated manner, using either a supervised (e.g., [21]) or unsupervised approach (e.g., [22]).

A lot of methods involve using other lexical resources, such as thesauri and lexicons, often taking the form of semantic networks. If a sufficiently large semantic network is available, in order to identify the polarity of words, a Markov random walk model can be applied (which generates a polarity estimate for a given word), as shown for instance in [22] or [23]. Moreover, many researchers (e.g., [21]) use web documents to compile polarity lexicons. Such lexicons are not restricted to any specific word classes and, in addition, contain slang and multi-word expressions. The experiments reveal that they allow for a vastly enhanced performance in polarity classification tasks when compared to lexicons based on, for example, WordNet.

In our work we attempt to create a polarity lexicon based on web documents. To the best of our knowledge, there exists no polarity lexicon for Polish that is freely available to the public and no research initiative seems to be aimed at compiling one. Yet, there are a number of general resources, such as Slowosiec (plWordNet)[1], the biggest Polish wordnet consisting of 94523 synsets, 133071 lexical units and almost 150000 lexical relations, which can be used in devel-

[1] http://nlp.pwr.wroc.pl/en/tools-and-resources/slowosiec

oping polarity lexical resources. Moreover, a set of tools put to use in language processing tasks (e.g., Morfologik, a morphological analyser [24]) prove helpful as well.

In contrast to the already existing (and otherwise interesting) research into identifying the sentiment value of texts written in Polish (e.g., [9, 25–27]), the authors of this work decided to adopt the lexicon-based approach to creating polarity lexical resources for this language. These are, moreover, intended to be subsequently made available to the general public.

There are a number of studies aimed at creating polarity (positive and negative) lexicons from corpora for languages other than English which make use of automated methods (e.g., [28,29]). Similarly to research based on such lexicons, our experiments show about 80% success rate with respect to identifying the polarity of adjectives and adjective phrases.

The section to follow is devoted to both discussing the mechanism used to create the above mentioned resources and presenting the assumptions underpinning the study.

3 Polarity Lexicon for the Polish Language: Fundamental Assumptions

The study sets out to devise a dedicated knowledge representation structure adequate for sentiment analysis. After a thorough examination of the available resources, the authors decided to create a semantic network capable of satisfying the needs of the conducted research, i.e., one with a core build in an automated manner and based on a number of freely accessible resources, such as dictionaries, thesauri and word lists taken from the existing open source projects, as well as web-based documents. The latter were gathered from portals that provide reviews on various goods and services. The total number of the collected reviews amounted to 356275. Apart from its content, each review was endowed with a score given by its author.

The structure of the semantic network is designed to store basic data on the type of relation (synonymous, hypernymous or homonymous) between individual concepts. In addition, each term has a set of extra properties, the key one being the sentiment vector holding data on the correlation of a given term with one of three basic sentiments (i.e., positive, negative and neutral) in a given domain.

In contrast to other previously made research efforts, the authors decided to extend the performed sentiment analysis to all parts of speech. If a concept was considered significant for a given domain, its sentiment score was stored in the semantic network together with information on the relevant domain. Accordingly, the polarity semantic network built up for the purpose of this study can be defined as follows:

The proposed polarity semantic network is a domain-aware sentiment lexicon Ls. Each element in Ls is associated with pairs (s_i, d_j)*, where* s_i *is the sentiment score for a given ith element in a domain* d_j*.*

In addition, the network contains data on the frequency of each concept, as culled from the reference corpus [30]. This appears to be a substantial improvement on the already existing semantic networks which have been observed to suffer from over-specializing when used for generalisation or summarising tasks. Providing frequency counts enables the algorithms used in the study to give a considerable advantage to more popular terms, thus making the generalised or abstracted text more easily accessible to a human user. While this may not be of critical importance in unsophisticated categorisation tasks, it is vital when evaluating various algorithms by humans.

In the course of the ongoing research on polarity lexicons for the Polish language an entirely new semantic lexicon was built. It stores data on over 140000 concepts. Each concept is not only described in terms of the semantic relations of hyponymy, hypernymy and synonymy but also accompanied by a sentiment vector providing data on its sentiment value in a given domain. Such extension was possible thanks to the availability of a specially collected corpus of documents with appended satisfaction scores. It was assumed that the documents with a high score contained words carrying positive connotations and those with a low score comprised negatively loaded terms. Understandably, the extraction of such words needed to be done with a sufficiently robust algorithm capable of disregarding words that have negligible influence on the sentiment of a whole document.

The corpus used in the study was based on the opinions provided by the users and accompanied by the overall product or service score. Thus, the end product the user was presented with was a concept with an attached vector expressing its polarity in specific recognisable domains. It was assumed that the most flexible approach to take would be to assign a normalised value to each concept. Normalisation was considered to involve mapping the available score on a scale of 0 to 10, where 0 stands for a maximally negative and 10 for a maximally positive sentiment value.

The words that successfully went through the verification process involving the use of the discussed algorithm were assigned a score value from a given document. The procedure was repeated on all of the available documents. Therefore, the sentiment score of an admitted word was the average of all the observed inputs. In addition to averaging the score, information on the domain a word came from was included to boost the postulated adaptivity.

Having accomplished that, it proved possible to test the effectiveness of the new resource on the documents that were excluded from the training phase. Even though the training corpus was of considerable size, the total number of concepts that were incorporated into the polarity lexicon constituted less 10% of the entire semantic network. Therefore, each document was represented by a surrogate that contained the basic forms of words deemed to carry some emotional load in a given domain.

The verification involved calculating the sentiment score using the polarity lexicon and, subsequently, comparing the results with the score provided by the author of a given document. The success rate of the verification process for the tested documents was close to 80%.

4 Random Walk: Extending Polarity Resources

As described above, the main strategy implemented to incorporate polarity data into the prepared semantic lexicon was the analysis of the available opinion corpus. It was observed that creative use of language prevents the existence of terms that can be only positive or only negative.

When faced with the data such as those used in the experiment, it needs to be born in mind that there exist structures in language, on the level of grammar and lexis alike, which are capable of producing the exact opposites of the meanings of lexical items. The most straightforward example would be a direct negation in the form of the word nie 'not' but less immediately obvious instances thereof were also found in the data gathered for the experiment. These included formulations such as daleki od (zadowalającego) 'far from (satisfactory)', mało (ciekawy) 'little (interesting)', trudno nazwać (pomocnym) 'difficult to call (helpful)' and constructions marking contrast (e.g., podczas gdy, chociaż/choć, wprawdzie, ale).

As proposed by [23] for Swedish, based on the efforts of [22], given an additional resource containing a set of words and their synonyms, along with a measure of the strength of synonymy according to the users, an alternative method can be implemented. This method provides a valuable input for a polarity lexicon, as it introduces a general notion of sentiment attached to a particular word. Such an extension offers additional benefits in terms of flexibility in a situation, where a domain remains unspecified or ambiguous.

To the best of our knowledge, the resources available for Polish are insufficient to employ any of the strategies proposed in the cited works. This is due to the fact that one cannot use a comprehensive thesaurus annotated with the data on perceived similarity between various synonyms. Although there are research initiatives aimed at providing comprehensive lists of synonyms with a division into senses, they all lack a quantifiable measure of determining how one synonym is related to another.

Therefore, the authors decided to devise a method in accordance with the rules formulated in [22], yet tailored to suit the resources developed in the course of the ongoing research activities.

The general algorithm is as follows:

- Start with a sentiment seed list
- Randomly choose a relation for the next move:
    - hypernymy (sentiment score is reduced by a factor of 0.3)
    - synonymy (sentiment score is reduced by a factor of 0.01)
    - hyponymy (sentiment score is reduced by a factor of 0.15)
- if a relation produces an element that was already used in the current walk, start over again
- if a relation produces more than one element, choose one at random and calculate a sentiment score

- store data on a visited element and its score
- if a relation produces no elements, start over
- if all nodes were visited or the limit of the repetitions of a random walk is reached, calculate the average sentiment score for terms that appeared more than 3 times in all random walks.

There were two research paths followed while testing the idea of extending the available polarity lexicon with the random walk algorithm. The first one involved using an initial seed of terms provided by the experiment participants. The whole list comprised 115 terms, along with the perceived sentiment scores abstracted from any domain. The number of terms obtained as a result of implementing the random walk amounted to 809.

The other research effort was concerned with reusing the already available experiment results, i.e., 27000 terms with the previously established sentiment scores. The total number of terms produced by the random walk algorithm with the seed list equalled 63539. It has to be emphasized that, due to the fact that the terms in this seed list were ascribed to a certain domain, it is difficult to say if the findings on the sentiment value of a term in a given domain can be automatically generalised to other domains.

Apart from conducting automated experiments using the sentiment test corpus (discussed in the subsequent section), the authors have also examined the result sets manually to assess their eligibility for inclusion in the polarity lexicon.

5 Evaluation of the Lexicon

The performed experiment used the already available test corpus. It contained 3222 documents, half of which were classified as having a positive and the other half a negative sentiment value. The experiment was carried out in the following stages:

- the documents were converted into surrogates,
- one of the three variants of the sentiment lexicon was used to prepare a prognosis about the sentiment of a document,
- the prognosis was compared with the actual score.

The results are summarised in Table 1.

Table 1. Results of the automated experiments using the sentiment test corpus

Sentiment lexicon	Valid for sentiment evaluation	Success rate
User provided seed	2763	61.49%
Extended original lexicon	3212	69.70%
Original lexicon	2426	78.93%

The research findings indicate that both lexicons extended by means of the random walk represent a marked improvement on the original experiments [3] in terms of determining whether a document is suitable for sentiment evaluation. As documents in the opinion corpus are rather brief, each term potentially influencing the sentiment value of the whole document is of crucial importance. It can also be observed that the success rate of measuring the sentiment of a document is considerably lower for the extended sentiment lexicons than for the original one.

Importantly, a small seed of user-provided terms with their general sentiment value proved to be highly effective, which is why future research is intended to focus on further activities aimed at building up this resource.

The lexicons extended by means of implementing the random walk were carefully examined for measures that can be taken to both retain a sufficiently large number of documents suitable for sentiment evaluation and improve the success rate of the entire evaluation process. Preliminary results of the analysis indicate that the structure of the underlying semantic network needs to be further refined. What is more, a heavier emphasis must be placed on adjectives, adverbs and nouns derived from adjectives.

6 Conclusions

The aim of the presented work was to demonstrate research efforts focused on preparing and expanding a polarity lexicon for the Polish language. Compiling such lexicon proved possible due to the availability of the previously built semantic network containing over 140000 concepts connected with each other by means of synonymic, hyperonymic and hyponymic relations.

It is the authors' intention to make the created polarity lexicon available to other researchers and the general public, once it contains a considerable number of concepts from diverse domains. At this point, the lexicon contains over 27000 concepts with a suggested sentiment value for a specific domain.

The experiments using the random walk algorithm and the available semantic network yielded more than twice as many candidates as the original tests, which is already a satisfactory outcome. Yet, without further improvements on the structure of the semantic network and implementing mechanisms allowing the inclusion of feedback from the users, the authors cannot fully guarantee the quality of this tool in terms of its ability to recognise and forecast the sentiment of a document (or its fragment) from an unrestricted domain.

Accordingly, future research initiatives are going to be devoted to preparing a set of software tools capable of delivering well-defined user-provided sentiment scores for individual concepts and phrases available in the created semantic network. What is more, it is envisaged that the tool kit will allow for both maintenance and expansion of the existing network.

As regards major structural changes, the underlying semantic network needs to be endowed with the notion of negation understood broadly as a grammatical and lexical phenomenon which may easily skew the results of sentiment analysis. Given the multiplicity of negation strategies and the effort that goes into producing the correct extraction rules for them, the authors initially strive to include just the most frequent ways of expressing negation, leaving the less immediately obvious ones to a later stage of research.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2(1-2), 1–135 (2008)
2. Tromp, E., Pechenizkiy, M.: Senticorr: Multilingual sentiment analysis of personal correspondence. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDMW 2011, pp. 1247–1250. IEEE Computer Society, Washington, DC (2011)
3. Haniewicz, K., Rutkowski, W., Adamczyk, M., Kaczmarek, M.: Towards the lexicon-based sentiment analysis of polish texts - polarity lexicon. In: Proceedings of ICCI. LNCS (LNAI). Springer, Heidelberg (2013)
4. Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., et al.: Memorandum on design-oriented information systems research. EJIS 20, 7–10 (2011)
5. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. Management Information Systems Quarterly 28(1), 75–106 (2004)
6. Hevner, A.R.: The three cycle view of design science research. SJIS 19(2), 87–92 (2007)
7. Gliwa, B., Kozlak, J., Zygmunt, A., Cetnarowicz, K.: Models of social groups in blogosphere based on information about comment addressees and sentiments. CoRR abs/1301.5201 (2013)
8. Zhang, C., Zeng, D., Li, J., Wang, F.Y., Zuo, W.: Sentiment analysis of chinese documents: From sentence to document level. J. Am. Soc. Inf. Sci. Technol. 60(12), 2474–2487 (2009)
9. Kowalska, K., Cai, D., Wade, S.: Sentiment analysis of polish texts. International Journal of Computer and Communication Engineering 1(1), 2010–3743 (2012) ISSN 2010-3743
10. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, vol. 10, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)
11. Paltoglou, G., Thelwall, M.: A study of Information Retrieval weighting schemes for sentiment analysis, pp. 1386–1395. Association for Computational Linguistics (July 2010)
12. Choi, Y., Cardie, C.: Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 590–598. Association for Computational Linguistics, Singapore (2009)
13. Sagot, B.: Introduction. In: Proceedings of WoLeR 2011, The 1st International Workshop on Lexical Resources (2011)
14. Maks, I., Vossen, P.: Different approaches to automatic polarity annotation at synset level. In: Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011 (2011)
15. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst. 21(4), 315–346 (2003)

16. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 347–356. ACM, New York (2011)
17. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417–422 (2006)
18. Strapparava, C., Valitutti, A.: WordNet-Affect: an affective extension of WordNet, vol. 4, pp. 1083–1086. Citeseer (2004)
19. Bradley, M.M., Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings. Psychology Technical (C-1) (1999)
20. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM 38, 39–41 (1995)
21. Velikovich, L., Blair-Goldensohn, S., Hannan, K., McDonald, R.: The viability of web-derived polarity lexicons. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 777–785. Association for Computational Linguistics, Los Angeles (2010)
22. Hassan, A., Radev, D.: Identifying text polarity using random walks. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 395–403. Association for Computational Linguistics, Stroudsburg (2010)
23. Rosell, M., Kann, V.: Constructing a swedish general purpose polarity lexicon: Random walks in the people's dictionary of synonyms. In: Proceedings of the Conference SLTC. Linkopings Univ., Sweden (2010)
24. Milkowski, M.: Automated Building of Error Corpora of Polish. In: Lewandowska-Tomaszczyk, B. (ed.) Corpus Linguistics, Computer Tools, and Applications - State of the Art, PALC 2007, pp. 631–639. Peter Lang, Frankfurt am Main (2008)
25. Wawer, A.: Extracting emotive patterns for languages with rich morphology. In: Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2012 (2012) (forthcoming)
26. Przepiórkowski, A.: A preliminary formalism for simultaneous rule-based tagging and partial parsing. In: Rehm, G., Witt, A., Lemnitzer, L. (eds.) Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007, pp. 81–90 (2007)
27. Buczynski, A., Wawer, A.: Shallow parsing in sentiment analysis of product reviews (2008)
28. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of html documents. Proceedings EMNLP-CoNLL (2007)
29. Maks, I., Vossen, P.: Building a fine-grained subjectivity lexicon from a web corpus. In: Calzolari, N., et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Istanbul (2012)
30. `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`

Identifying Discriminatory Characteristics Location in an Iris Template

Alaa Hilal[1,2,*], Pierre Beauseroy[1], and Bassam Daya[2]

[1] UMR STMR – LM2S – ICD, Université de Technologie de Troyes, Troyes, 10004, France
{alaa.hilal,pierre.beauseroy}@utt.fr
[2] Université Libanaise, Beyrouth, Liban
b_daya@ul.edu.lb

Abstract. The high stability, uniqueness and richness of the iris make it among the best biometric templates. However while examining it one can find that different iris sub-regions result in heterogeneous recognition performances. In consequence, recognition decisions based from partial regions of the iris are subject to the regions themselves. In this paper we investigate the location of discriminatory characteristics in the iris by evaluating recognition performance drawn from different concentric rings within the iris. Iris images are first segmented using Hough transform and active contour. Then a robust normalization takes into account the pupil's free shape and its center of gravity to unwrap the iris. Gabor filter and Hamming distance are next applied for the encoding and the matching respectively. Results obtained, on 2491 iris images from CASIA-V3 database, show that, contrary to previous studies, the most discriminating characteristics are located in the inner regions of the iris.

Keywords: Biometrics, Image Processing, Iris Characteristics Location, Iris Recognition, Hough Transform, Active Contour.

1 Introduction

Human iris template is generated by a number of accumulated layers that result in a heterogeneous texture containing fibers, contraction furrows, crypts, rings and freckles. These elements, characterized by their randomness, uniqueness and their stability, make the iris among the best biometric templates. Consequently an iris recognition system that makes use of the iris template can be developed to identify an unknown identity. The system takes an image of the unknown iris and makes use of a segmentation procedure to extract the iris region from the image, a normalization method to transform the iris into a fixed-size template, and finally an encoding process in order to extract from the iris its discriminatory characteristics. When identifying the unknown iris, the processed iris code is compared to other codes from recognized iris code database in order to reveal the unknown identity. High

[*] Corresponding author.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_18, © Springer International Publishing Switzerland 2014

recognition performance results are generally reported validating the discriminatory characteristic of the iris [1], [2].

On the other hand, when observing any iris image it is noticeable that the density of the texture is not homogeneous all over the iris. Generally it appears that dense texture is more located in the inner and in the middle regions rather than the outer regions of the iris template. Consequently one can hypothesis that recognition performance varies when considering specific sub-regions of the iris. Such fact can be used to asses a recognition decision in cases like when only partial parts of the iris are only considered for the recognition. Thereby developing a better knowledge of how discriminatory characteristics are located all over the iris region is of major advantage.

In this paper, we investigate iris information location by measuring various recognition performance results obtained when selecting different regions of the iris template. We start our paper, by reviewing the related work in section 2. Next we introduce the developed iris recognition system we used for the search of iris discriminatory characteristics location in section 3. Obtained results and conclusions are finally detailed in section 4 and 5 respectively.

2 Related Work

Limited studies and experiments have been reported on the location of iris's discriminatory characteristics. All of them use approximately same approaches by considering different regions of the iris made by decomposing the iris into different concentric rings. Recognition performance results are then measured when considering each ring separately for the identification process. When building the iris recognition system, methods that are similar to Daugman's system have been used [2]. Both iris boundaries are firstly approximated by circular or elliptical contours. Segmented iris is then normalized into a rectangular matrix following Daugman's "rubber sheet" model. The obtained template is encoded and matched to enrolled iris signatures finally.

Results obtained, can be classified into two categories. A first group of researchers states that discriminatory iris characteristics in iris regions that are close to the pupil [3], [4], while a second group of researchers supports the hypothesis that discriminatory iris characteristics exist in the center rings of the iris [5-7].

In more details, Du et al. segmented the iris using circular approximation methods. The iris is then normalized similar to Daugman's method by remapping the segmented template into a rectangular matrix using concentric circles that cover the iris region. Average local intensity variance is then calculated to create the iris code (signature). Matching iris codes is then processed by calculating a similarity measurement between registered and unknown codes. This similarity metric called "Du measurement" is based on three operators that take into consideration relative entropy variation, angle variation, and average power difference between the signatures. The study was tested on 742 images from CASIA V1 database and on 818 images from another database. Results they obtained showed that discriminatory characteristics are found in the inner rings of the iris and attenuates with rings that are close to the limbic boundary [3], [4].

Gentile et al. used ellipse fitting to detect iris boundaries. Normalization is done similar to Daugman's method. However instead of using concentric circles to remap the iris into a rectangular fixed-size matrix, elliptical contours are virtually drawn all over the iris. Center of these ellipses are defined on the segment defined by the pupil's limbic elliptical contours centers. 1D log-Gabor filter and Hamming distance was then used for encoding and the matching of the iris. Results are obtained from 425 iris images of MMU2 iris database images [5].

Broussard et al. segmented and normalized the iris following Daugman's methods. However they used directional energy obtained from a cosine kernel filter response to encode the iris and Hamming distance for the matching purpose. The study is tested on 367 iris images from the University of Bath database [6].

Hollingsworth et al. segmented and normalized the iris following Daugman's methods also. To encode the iris they used the quantized phase response of 1D log-Gabor filter response. Matching iris signatures was later performed by measuring the Hamming Distance. The study was performed on 1226 images from Challenge Evaluation database [7].

The three groups Gentile et al., Broussard et al. et Hollingsworth et al. concluded that iris information are more discriminate close to the center rings of the iris, whereas iris regions that are close to the pupil or to the sclera present a less discriminatory capability [5-7].

3 Developed Iris Recognition System

In order to investigate the iris discriminatory characteristics location, a well developed iris recognition system must be used. Otherwise the results accuracy may be decreased by weak methods used in the system. When evaluating the previous work we can observe that:

- Contradictory results have been found on the location of iris information; a first group concludes that inner regions contain the most discriminative information [3], [4], whereas the center regions were the most discriminative information in an iris following the results of a second group [5-7].
- A greater number of textural structures is reported in the inner regions of the iris but even thought they resulted in a weak discrimination capability [5-7].
- All of the previous work did not use a precise segmentation method. Instead they used circular or elliptic approximation methods. While this approach may fit the limbic boundary of the iris, it is not the same case for the pupil's boundary. As for the normalization, Daugman's approach was implemented. Normalization strips were either concentric circles, or circles or ellipses whose centers are defined on the segment joining the pupil's center and the limbic boundary center [3-7].
- The presence of less discriminatory information in inner regions of the iris can be related with pupil's dilation, and the inaccurate segmentation and normalization [5-7] of the iris. As for the outer regions of the iris, the weak recognition performance was related to the lack of iris structure [5].

Based on these facts, one can interpret that segmentation and normalization of the iris is of great impact on recognition performance. Taking into account all of the previous facts, we propose to search the location of discriminatory characteristics of the iris following a developed iris recognition system. Iris segmentation is done using a circular approximation of the limbic boundary and a free shape detection of the pupil's boundary. Then normalization of the iris is done using elastic normalization strips that remap the iris into a rectangular matrix. However the elastic normalization strips varies from the pupil's free shape into the limbic circular contour while normalizing the iris. As for the center of these strips, it varies on the segment joining the center of gravity of the pupil and the center of the circle approximating the limbic boundary. Encoding and Matching are finally performed following Daugman's system that uses Gabor filters and Hamming distance measurement respectively. The developed system is explained in what follows.

3.1 Segmentation Using Hough Transform and Active Contour without Edges

Acquired iris images are firstly segmented. The process refers to locate the iris and isolate it from surrounding noise such that eyelids, eyelashes and specular reflections. Methods developed to locate the iris can be grouped into two categories: circular or elliptic approximation and model free contour detection. In the first category iris boundaries are detected with circular or elliptical fitting operators such as methods used with Daugman [2] and Wildes [8]. The second category of segmentation methods uses free contour segmentation as active contour methods with Daugman [9] and with Shah and Ross [10].

The model of segmentation we've developed makes use of circular Hough transform and active contour without edges. The limbic boundary is firstly approximated with a circular contour and then pupil's boundary is detected in a free shape contour [11], [12].

Hough transform is an object recognition operator. It detects objects with simple shapes such as lines and circles. A binary edge map is calculated from the iris image using a gradient filter. From the edge points obtained, votes in a Hough space are counted to estimate the circle's parameters: center (x0,y0) and radius (r) that approximate the iris boundary [8].

Active contour without edges is an iterative segmentation method introduced by Chan and Vese [13]. Starting with an initial contour, the active contour evolves iteratively toward object contour in an image. Contour evolution is defined by an energy function that is based on region information rather gradient information. Considering an image with intensity value I(x,y) the each pixel (x,y), we designate by C an object's contour inside the image. Let c1 and c2 be the average intensities inside and outside C respectively. The active contour energy is defined as:

$$E(c_1, c_2, C) = \sum_{Interior(C)} |I(x,y) - c_1|^2 + \sum_{Exterior(C)} |I(x,y) - c_2|^2 \quad (1)$$

The convergence of the contour is assured by minimizing the function E in terms of c1, c2, and C enables the active contour to converge to the contour of the object. The displacement of the contour from iteration to another is given by the equation introduced by Chan and Vese [13]:

$$\frac{\partial C(x,y)}{\partial t} = div\left(\frac{\nabla C(x,y)}{|C(x,y)|}\right) - (I(x,y) - c_1)^2 + (I(x,y) - c_2)^2 \tag{2}$$

where $C(x,y)$ represent the intensity of the set of points that define the contour C.

To apply our segmentation, Hough transform is firstly applied to detect the limbic boundary, and then applied once again to approximate the pupil's contour with a circle. Result obtained is used as an initial contour in order to evolve the active contour into a precise detection of the pupil's boundary. Once both boundaries are detected, eyelids were isolated by applying a linear Hough transform to approximate the eyelid shape. Eyelashes and specular reflections are finally eliminated by intensity threshold operations.

3.2 Elastic Normalization

Once segmented, iris region is to be transformed into a fix-sized template. We use in our system a developed normalization method that takes into account the shape free model of the pupil's contour. Instead of remapping the iris using circular or elliptical strings, we use elastic strings with a shape that varies between the pupil's free shape and the circular approximation of the iris. Considering the segmented iris region is characterized by an intensity value $I_C(x,y)$ in each pixel (x,y) in the Cartesian space, the newly normalized iris intensity $I_P(r,\theta)$ in each pixel(r,θ) in the Polar space is given by the following equations:

$$I_P(r,\theta) = I_C\big(x(r,\theta), y(r,\theta)\big) \tag{3}$$

with

$$x(r,\theta) = (1-r)x_p(\theta) + rx_s(\theta) \tag{4}$$

$$y(r,\theta) = (1-r)y_p(\theta) + ry_s(\theta) \tag{5}$$

$\left(x_p(\theta), y_p(\theta)\right)$ and $\left(x_s(\theta), y_s(\theta)\right)$ are the coordinates of the pupil and limbic boundaries respectively at angle $\theta \in [0,2\pi]$, r varies between 0 and 1 which corresponds to the pupil and limbic boundaries respectively.

Centers of the normalization strips for each radius value, r, move on the segment defined by the pupil's center of gravity and the center of circle approximating the limbic boundary center while unwrapping the iris starting from the pupil to the sclera.

A binary mask that marks noisy pixels in the iris template is generated. It is used to ignore noisy pixels while matching two iris codes.

When comparing our method to Daugman's normalization, a major difference is that normalization strips used to unwrap the iris. These strips are always circular and concentric in Daugman's method. However in our system they vary between the pupil's accurate contour and the limbic boundary approximation.

After normalization, iris templates are encoded and matched. Daugman's approaches were used. Encoding process refers to extract from the iris its most significant characteristics. 2D Gabor filter given by the equation below is employed:

$$G(r,\theta) = e^{-iw(\theta)}e^{-(r)^2/\alpha^2}e^{-i(\theta)^2/\beta^2} \tag{6}$$

where (α,β) are the effective width and length and ω is the filter's angular frequency. The phase response of the filter is quantized with two bits to represent the iris signature.Obtained iris codes are matched using the normalized Hamming distance (HD). The method measures the correspondence similarity between the two iris codes. For two different iris codes A and B having binary noise masks Amask and Bmask respectively, the Hamming distance between A and B is given by:

$$HD(A,B) = min_{\varphi}\left\{\frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left\{|A(i+\varphi,j)-B(i,j)|\times\left(A^{mask}(i+\varphi,j)\times B^{mask}(i,j)\right)\right\}}{\sum_{i=0}^{M}\sum_{j=0}^{N}\left(A^{mask}(i+\varphi,j)\times B^{mask}(i,j)\right)}\right\} \tag{7}$$

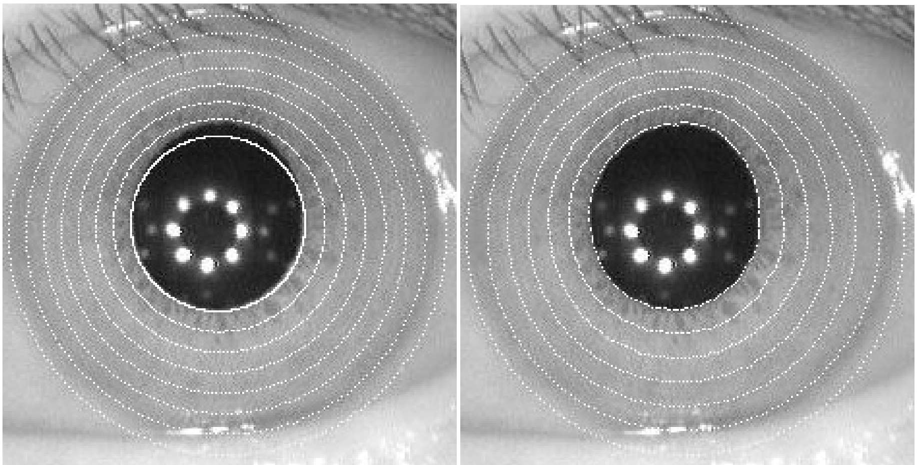
for $-10 \leq \varphi \leq +10$ where φ represents a scaling parameter that compensates iris rotation, *M* and *N* are respectively the angular and radial size of the iris code, φ is a translational parameter that compensates iris rotation. Iris codes generated from a same iris will result in a low *HD* value (close to 0) whereas when the codes are generated from two different irises *HD* will have higher *HD* value (close to 1) [2], [9].

Following our proposed iris recognition system, each iris region is divided into 10 concentric rings. Each ring will be then processed separately in order to measure recognition performance and assess its discriminatory capability.

4 Results

The developed system is applied on CASIA V3-Interval iris images database. *2491* images are used to identify the location of discriminatory characteristics in the iris [14]. Segmented iris images are normalized to a dimension of 40 × 240 pixels. Each iris is then divided into 10 concentric rings that correspond to rectangular sub-matrixes in the normalized iris. Each ring corresponding to a sub-matrix will have a dimension of 4 × 240 pixels.

In order to evaluate the discriminatory capabilities of different regions of the iris, recognition performances are measured from considering separately each ring of the iris templates. Obtained results are also compared to those obtained when using Daugman's system [2]. In this system iris images are segmented following a circular approximation of iris boundaries using Daugman's integro-differential operator. Segmented images are next normalized according to Daugman's "rubber sheet" model. Then 2D Gabor filters are used for iris encoding and Hamming distance is used for the matching purpose. Difference in Daugman's normalization and our elastic normalization is presented in Figure 1.



(a) (b)

Fig. 1. Iris normalization according to Daugman's method (a) and to our proposed method (b) White dashed contours are the normalization strips applied in each method

The white dashed contours in both images show the normalization strips in each method. Normalization strips are circular with Daugman's method however they do not cover accurately the pupil's boundary. In our method, the elastic normalization covers accurately the iris region between its two boundaries.

Recognition performance drawn from considering the ten rings of iris images separately is measured using the two parameters: decidability, accuracy at Equal Error Rate (EER).

4.1 Decidability

Decidability is a performance parameter that measures the separation between the intra and the inter-class *HD* scores obtained when evaluating two iris codes from a same class or from different classes respectively. Let (μ_s, μ_D) and (σ_s, σ_D) be the mean and the standard deviation of the intra and inter-class Hamming distance scores respectively [2]. The decidability can be measured according to the following equation:

$$Decidability = \frac{|\mu_S - \mu_D|}{\sqrt{(\sigma_S^2 + \sigma_D^2)/2}} \qquad (8)$$

Higher decidability value reflects a more important separation between the intra and the inter-class scores. The decidability results obtained from applying our developed iris recognition system and from applying Daugman's system to each of the 10 rings of all iris images are drawn in figure 2.

It appears that following Daugman's system, rings 3 to 6 hold the highest decidability values between all the rings and hence they reflect the highest discriminating power among the other rings.

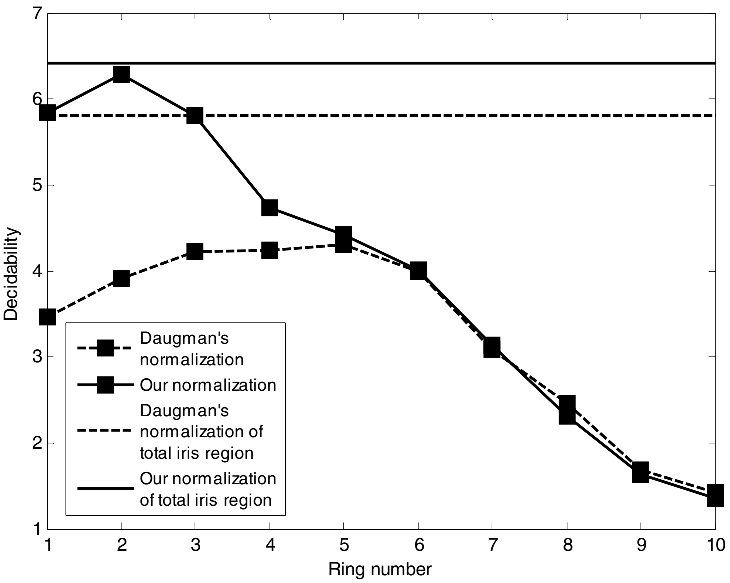


Fig. 2. Decidability of each of the ten rings of the iris following the two methods (Ring 1 is the closest to the pupil; ring 10 is the closest to the sclera)

Lower decidability values are then obtained with rings 1 and 2 whereas the lowest values ever are obtained with rings 7 to 10 that are the closest to the sclera. These results confirm the conclusions obtained with previous work [5-7].

However when considering our developed iris recognition system results, we can notice that the highest decidability values are obtained for the first inner three rings of the iris. Even more the decidability values of these three rings are equal or greater than that obtained when considering the entire iris using Daugman's system. Starting from ring 4 to the last ring of the iris, the decidability values are almost identical to those resulted from Daugman's system.

4.2 Accuracy at Equal Error Rate

A second parameter we used to assess recognition performance is the accuracy at equal error rate (EER). Considering two iris codes to be matched, the *HD* between these two codes is calculated. A match is obtained when the measured *HD* is lower than a threshold decision. Two types of recognition errors can be obtained: false positive and false negative recognition. False positive occurs when a match decision between two iris codes is obtained given that the two codes do not correspond to the same iris. False negative is obtained when a non match between two iris codes is obtained given the two codes do correspond to a same iris. Accumulating the false positive and the false negative errors over the tested images results in the calculation of the false accept rate (FAR) and the false reject rate (FRR) respectively [2].

ERR corresponds to the rate of error when FAR is equal to FRR. The accuracy of recognition at EER is used as a measure of system's performance. Figure 3 represents the accuracy at the EER obtained for each of the ten rings of the iris using the two tested iris recognition systems.

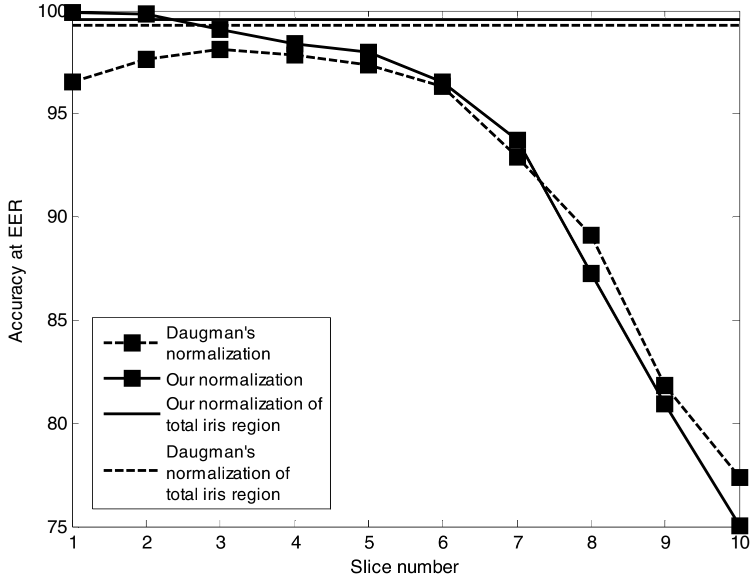


Fig. 3. Accuracy at the EER obtained with each of the ten rings of the iris following Daugman's and our approach (Ring 1 is the closest to the pupil; ring 10 is the closest to the sclera)

Considering Daugman's system, best values of accuracy at EER are obtained for the middle rings of the iris (rings 3 to 5). This accuracy drops from nearly 98 % in the middle rings to a 96 % in the inner rings of the iris. The accuracy at EER decreases dramatically with the outer rings to reach a value of 75 %.

These results aren't the same with our developed iris recognition system. In fact the best accuracy recognition rates are obtained for inner rings of the iris in our system. Rings 1 to 3 have an accuracy value at EER that is greater than 99.5%. This value decreases in parabolic form starting from ring 4 to the last ring of the iris. Accuracy values in these rings take nearly the same values as with Daugman's system. Finally we can observe that accuracy at EER values for rings 1 and 2 in our iris recognition system are higher than Daugman's system performance when using the entire iris region.

The coherent decidability and accuracy at EER results show that inner regions are more performing than the middle or the outer regions of the iris in the recognition process. These results demonstrate that inner regions hold the most discriminatory characteristics in an iris template. Also as one traverses the iris from the pupil's boundary to the limbic (outer) boundary, the discriminatory characteristics of the iris template decreases.

5 Conclusions

Accurate knowledge of the location of discriminatory characteristics in an iris template is of high importance. Knowing whether iris information are more discriminate in inner, middle or outer regions of the iris reflects how much a recognition decision that is based on a partial region of an iris is trustable or not. In non-ideal iris images case, it could be thus sufficient to only consider a portion of the iris for the recognition process.

When evaluating the previously reported work, contradictory results argue whether iris information is more discriminate in the inner or in the middle regions of the iris. However no accurate segmentation of the pupil's boundary neither accurate normalization of the inner regions of the iris were followed in the reported work.

In our work we build our hypothesis that inner regions of the iris hold the most discriminating characteristics. We developed our iris recognition system to accurately segment the iris boundaries and normalize the iris template. The obtained iris template is then divided into 10 rectangular templates representing 10 concentric rings of the iris. Our developed iris recognition system was then compared with Daugman's system. Recognition performances of the two systems from using each of the 10 rings separately are drawn. The decidability and accuracy at EER recognition performance parameters showed concordant results.

When using our developed system, the most discriminate characteristics in an iris template are found in inner regions of the iris that are the most close to the pupil. Starting from these regions the discriminate potential of characteristics decreases to reach its minimum in regions that are close to the limbic boundary. However when using Daugman's system it appears that middle and outer regions of the iris presented performance results that are similar to ours. However a drop in recognition performance is always observed in the inner regions of the iris.

We conclude that it was the weak segmentation of the pupil's boundary and the non-optimized normalization that resulted in weak recognition performance in the inner regions of iris templates in Daugman's system and in the previously reported work [5-7]; hence the false assumption of weakness of discriminatory characteristics in the inner regions of the iris. By improving the segmentation and the normalization of the iris we demonstrated that inner regions hold the most discriminating characteristics in an iris template.

References

1. Krichen, E.: Reconnaissance des personnes par l'iris en mode dégradé. Ph.D. dissertation, Evry-Val Essonne University (2007)
2. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1148–1161 (1993)
3. Du, Y., Bonney, B., Ives, R., Etter, D., Schultz, R.: Analysis of partial iris recognition using a 1D approach. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 961–964 (2005)

4. Du, Y., Ives, R., Bonney, B., Etter, D.: Analysis of partial iris recognition. In: Proceedings of Biometric Technology for Human Identification II, vol. 5779(31) (2005)
5. Gentile, J., Ratha, N., Connell, J.: SLIC: Short-Length Iris Code. In: IEEE International Conference on Biometrics: Theory, Applications, and Systems, Washington, pp. 1–5 (2009)
6. Broussard, R.P., Kennell, L.R., Ives, R.: Identifying discriminatory information content within the iris. In: Proceedings of Biometric Technology for Human Identification, Orlando (2008)
7. Hollingsworth, K.P., Bowyer, K.W., Flynn, P.J.: The Best Bits in an Iris Code. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), 964–973 (2009)
8. Wildes, R.P.: Iris recognition: an emerging biometric technology. Proceedings of the IEEE 85(9), 1348–1363 (1997)
9. Daugman, J.: New Methods in Iris Recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 37(5), 1167–1175 (2007)
10. Shah, S., Ross, A.: Iris Segmentation Using Geodesic Active Contours. IEEE Transactions on Information Forensics and Security 4(4), 824–836 (2009)
11. Hilal, A., Daya, B., Beauseroy, P.: Hough Transform and Active Contour for Enhanced Iris Segmentation. International Journal on Computer Science Issues 9(6) (2012)
12. Hilal, A., Beauseroy, P., Daya, B.: Real shape inner iris boundary segmentation using active contour without edges. In: International Conference on Audio, Language and Image Processing, Shanghai, pp. 14–19 (2012)
13. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing 10(2), 266–277 (2001)
14. CASIA-IrisV3, `http://www.cbsr.ia.ac.cn/IrisDatabase`

Background Modeling in Video Sequences

Piotr Graszka

Warsaw University of Technology, Warsaw, Poland
P.Graszka@ii.pw.edu.pl

Abstract. The background modeling is the very first and essential part of every computer assisted surveillance system. Without it there would be no reliable way for fast and robust detection of moving objects in video sequences. In this paper we collect, describe and compare the main features of the most commonly used techniques of background modeling in video sequences and determine the most desirable way for the development of new algorithms in this field.

Keywords: background modeling, background subtraction, image processing.

1 Introduction

1.1 The Description and Occurrence of the Problem

The video surveillance systems are currently widely used in different aspects of the daily life. The simplest ones are set up in almost every shop. More sophisticated installations are used for traffic observation or passenger behavior analysis at the airports. They often consist of many video cameras and recording devices thus observation and interpretation of the recorded material requires a correlated work of properly trained personnel. In order to improve the detection of specific types of situations or behaviors, computer assisted surveillance systems are used. Such setups are capable not only of detecting certain situations but also of conducting the initial analysis, so that oversight probability falls almost to zero.

1.2 What "The Background" Is and How to "Model" It

All computer assisted surveillance systems have one common feature: their work rely on motion detection algorithms. The segmentation technique that is most frequently used to detect motion in video sequences is the background subtraction. It owes its popularity mainly to the achievable performance which involves the possibility of real-time processing [2], [3]. The "background" is interpreted as a set of pixels that have constant over time properties, like for example the color or the frequency of intensity changes. However, to remove it, it is necessary to know how it looks like. This is where the background modeling methods come in as they are able to determine how the background looks. Having created the background model, it is easily determinable which areas of the image may contain interesting information, and that is the first and essential step in any automatic detection system [1-3].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_19, © Springer International Publishing Switzerland 2014

2 Background Modeling Methods

A good background modeling method should be able to adapt to three fundamental scenery changes. These are [1], [3]:

- change of the brightness, such as the sun coming out from behind the clouds,
- continuously repeating changes, such as a flag flapping in the wind,
- change in the geometry, such as a car driving away from a parking space.

In addition, the necessary condition in most practical applications is the possibility to work in the real time. Otherwise, the algorithm cannot be used for live streams.

In the following parts of this paper we characterize the most common approaches to the problem. All the tests were conducted using many different video sequences [12], but due to limited space all the illustrations in this paper present the outcome of described methods always using the frame shown in figure 1.

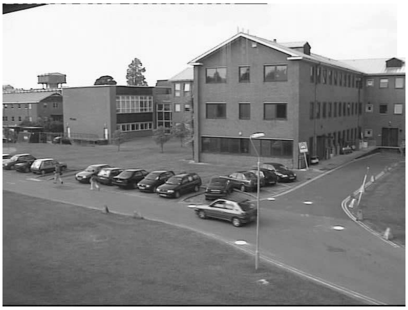


Fig. 1. Frame chosen to present the results returned by the described algorithms

The only moving objects in figure 1 are two people in the left part of the image and a car in the center. The ideal result of the background modeling in this case is a completely black image with three white areas corresponding to the moving objects.

2.1 The Difference of Subsequent Frames in a Video

The easiest way to detect motion areas is to check where the biggest differences between consecutive frames are. Mathematically, this relationship can be written as [1]:

$$| F(x,y)_n - F(x,y)_{n-1} | > T \tag{1}$$

where:
$F(x,y)_n$ – pixel from the n-th frame,
$F(x,y)_{n-1}$ – pixel from the (n-1)-th frame,
T – threshold value for movement occurrence.

This method is computationally very simple, thus also very fast. Unfortunately, the correct results are achievable only in specific circumstances, for example it is not possible to detect very slow movement because it introduces only minor changes in two consecutive frames. Moreover, the detection results are very sensitive to changes in the threshold value [1]. Too low threshold introduces a noise, while too high will prevent detecting a part of the moving objects.

In the paper [11] an improvement to the described approach was proposed: before the area is considered to be part of the background it must satisfy the condition (1) for a given number of frames. In this case the movement occurrence condition is [11]:

$$| F(x,y)_n - B(x,y)_{n-1} |> T \tag{2}$$

where:
$F(x,y)_n$ – pixel from the n-th frame,
$B(x,y)_{n-1}$ – background pixel estimated in (n-1)-th step,
T – threshold value for movement occurrence.

This approach works better, but still it does not cope very well with the varying background. In this case, the noise increases as well as ghost objects start to appear.

The sample results returned by this method are shown in figure 2. In addition to moving objects, a large amount of spot noise, which over time gets strengthened, can be seen in the picture. A positive feature of the illustrated approach is the very high processing speed, reaching more than 200 frames per second.

2.2 The Background as the Average of N Previous Frames

Much better results are provided by the methods that use statistical measures to estimate the background. The simplest one of them calculates the background as the arithmetic mean of the N previous frames [1]. To identify the foreground areas, a standard procedure of subtracting the estimated background from the current frame and comparing the result to a specified threshold is used. This approach is not computationally complex, so the algorithm runs fast, but has high memory requirements, because every time all N frames are needed for the calculations.

In the papers [1], [2], [5] an improvement to the described method was proposed which introduces a formula for a good approximation of the mean:

$$B(x,y)_n = \alpha \cdot F(x,y)_n + (1-\alpha) \cdot B(x,y)_{n-1} \tag{3}$$

where:
$B(x,y)_n$ – currently estimated mean,
$B(x,y)_{n-1}$ – previously estimated mean,
$F(x,y)_n$ – pixel from the n-th frame,
α – estimation factor, usually $\alpha \in (0.01, 0.1)$.

The algorithm works only a little faster, but the memory requirements become almost negligible in comparison with the previous one. The averaging of frames over time reduces the spot noise and a continuous actualization makes it possible to model a slowly varying background. However, the algorithm has some problems with a rapidly varying background, particularly if these changes have high variance. In addition, the wrong choice of the α parameter may cause the trails to remain behind the moving objects or an incorporation of slow movement to the background model.

A common way to prevent these types of errors is to update only the still areas – a selective update. This changes the background estimation formula to [1], [5]:

$$B(x,y)_n = \begin{cases} B(x,y)_{n-1} & \text{for (2)} \\ \alpha \cdot F(x,y)_n + (1-\alpha) \cdot B(x,y)_{n-1} & \text{otherwise} \end{cases} \quad (4)$$

This approach allows the detection of slowly moving or even standing still foreground objects. Unfortunately, it can also cause additional noise or ghost objects to form.

The sample results of the algorithm are shown in figure 3. It shows long trails remaining behind the moving objects. This error results from too rapid actualization causing a contamination of the model. In addition, a lot of spot noise can be seen that, in contrast to the simple differential method, does not grow over time.

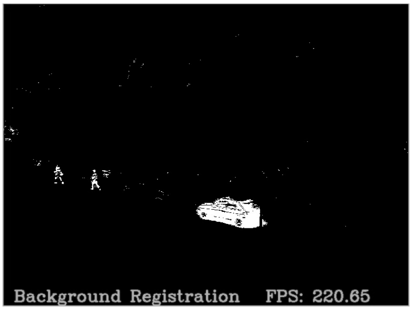


Fig. 2. Results returned by the frame difference with background registration method

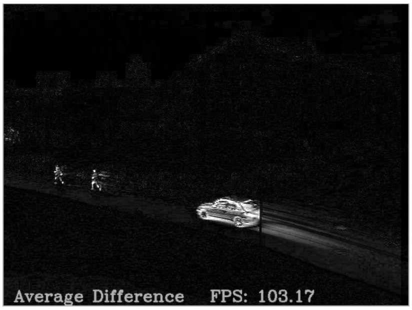


Fig. 3. Results returned by the average difference method

2.3 The Background as a Median of N Previous Frames

The median is another statistical measure often used in the context of background modeling. The basic idea is the same as for the mean [1], [3], [4], [6] and also the results are similar. However, the median method is more accurate – less noise and smaller trails remaining behind the moving objects. This is because random deviations from the majority of the analyzed data have very little effect on the median.

Paper [3] describes two major improvements that enable the approximation of the median. The first one uses the following recursive update function:

$$B(x,y)_{n+1} = \begin{cases} B(x,y)_n + 1 & \text{for } F(x,y)_n > B(x,y)_n \\ B(x,y)_n - 1 & \text{for } F(x,y)_n < B(x,y)_n \\ B(x,y)_n & \text{for } F(x,y)_n = B(x,y)_n \end{cases} \quad (5)$$

where:
$B(x,y)_{n+1}$ – currently approximated median,
$B(x,y)_n$ – previously approximated median,
$F(x,y)_n$ – pixel from the n-th frame.

Thanks to this improvement, the algorithm has very low computational complexity and memory requirements, thus it runs very fast not consuming a lot of resources.

The second improvement makes the method consider all three color channels of the input image together by calculating a three-dimensional vector median using formula:

$$\mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}+1} = \begin{cases} \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}} + \frac{\mathbf{F}(\mathrm{x},\mathrm{y})_{\mathrm{n}} - \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}}}{\|\mathbf{F}(\mathrm{x},\mathrm{y})_{\mathrm{n}} - \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}}\|} & \text{for } \mathbf{F}(\mathrm{x},\mathrm{y})_{\mathrm{n}} \neq \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}} \\ \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}} & \text{for } \mathbf{F}(\mathrm{x},\mathrm{y})_{\mathrm{n}} = \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}} \end{cases} \tag{6}$$

The movement occurrence condition has changed accordingly:

$$\| \mathbf{F}(\mathrm{x},\mathrm{y})_{\mathrm{n}} - \mathbf{B}(\mathrm{x},\mathrm{y})_{\mathrm{n}} \| > \mathrm{T} \tag{7}$$

The main advantage of this three-dimensional approach over the previously described methods is a full utilization of color information, so that the statistical relationships between the color components are taken into account in the background model [3].

The sample results returned by the algorithm using the first median approximation method are shown in figure 4 and are almost identical to those returned by computing the median directly from previous N frames. However, the approximated median algorithm runs much faster because the whole computing course is approximately 4 times shorter thanks to the simplifications used in the calculations.

2.4 The Analysis of a Histogram of N Previous Frames

A histogram is a statistical record of a number of consecutive values occurrences in a given data set. In case of the background modeling, the histogram is created over time for each pixel separately, so that it shows the distribution of brightness values over time. If it is interpreted as probability density distribution then it can be used to determine the occurrence probability of a new pixel. Then if the probability exceeds the specified threshold the new pixel is considered to be a part of the background and can be used to update the histogram; otherwise it belongs to the area of movement.

The algorithms that use this approach must overcome two serious problems:

- histogram discontinuity – it may happen that two adjacent values have radically different occurrence likelihood or that between two high probabilities there is a zero probability gap. Such situations are caused by a limited size and a discrete form of the data set from which the histogram was created. The most common way to solve this problem is blurring the values onto the adjacent intervals, which, to some extent, corrects the continuity of the resulting histogram.
- movement occurrence threshold selection – there are two commonly used solutions for this problem. The first one is to set the threshold to an empirically chose single constant value. Unfortunately it reduces the algorithm robustness to large changes in brightness, causing more noise to appear in such a case. The second solution requires the threshold value to be set and usually updated during the algorithm run based on the variance of the modeled background, which unfortunately cannot always be calculated [3].

Due to this problems and large memory requirements there are no practical methods that use the histogram directly to estimate the background. However, a number of methods utilize the advantages of the histogram by approximating it. The most common of them use continuous random distribution functions to meet that purpose.

2.5 The Histogram Estimation with the Gaussian Distribution

There are three most popular methods for the histogram estimation with the Gaussian distribution that result from one another [3]. The first of them is trying to render the shape of the histogram using a single time-varying Gaussian distribution described by the mean (μ) and the standard deviation (σ) with the following formulas [1]:

$$\mu(x,y)_{n+1} = \alpha \cdot F(x,y)_n + (1-\alpha) \cdot \mu(x,y)_n \tag{8}$$

$$\sigma(x,y)^2_{n+1} = \alpha \cdot (F(x,y)_n - \mu(x,y)_n)^2 + (1-\alpha) \cdot \sigma(x,y)^2_n \tag{9}$$

where:
$\mu(x,y)_{n+1}$ – currently approximated mean,
$\mu(x,y)_n$ – previously approximated mean,
$\sigma(x,y)^2_{n+1}$ – currently approximated variance,
$\sigma(x,y)^2_n$ – previously approximated variance,
$F(x,y)_n$ – pixel from the n-th frame,
α – estimation factor, usually $\alpha \in (0.01, 0.1)$.

The analyzed area is classified as a movement if it fulfills the condition [1]:

$$| F(x,y)_n - \mu(x,y)_n |> T \tag{10}$$

where:
$F(x,y)_n$ – pixel from the n-th frame,
$\mu(x,y)_n$ – previously approximated mean,
T – part of the standard deviation – kσ(x,y).

Thanks to such a form of the movement condition, the selectivity can be used in a simple manner. It is enough to update only those areas that belong to the background.

The method operates at a speed comparable to the average difference method. It also has a low memory requirements, because it has to remember only two values (μ, σ) for each pixel. Unfortunately, it has a serious drawback compared to the approach using the histogram directly. It is unable to model a rapidly varying background.

The second histogram estimation approach, the Gaussian Mixture Method, can cope with the rapidly changing background through the use of several Gaussian distributions, each of which has its own mean (μ), standard deviation (σ) and weight (ω) that indicates its usefulness for the background regions detection. This approach uses formulas (8) and (9) to approximate the mean and the standard deviation, but they are used only to update the distributions fulfilling the condition [7], [8]:

$$| F(x,y)_n - \mu(x,y)_{n,i} |< 2.5 \cdot \sigma(x,y)_{n,i} \tag{11}$$

where:
$F(x,y)_n$ – pixel from the n-th frame,
$\mu(x,y)_{n,i}$ – mean value of the i-th distribution,
$\sigma(x,y)_{n,i}$ – standard deviation of the i-th distribution.

The weight of each distribution is updated according to the formula [7], [8]:

$$\omega(x,y)_{n+1,i} = \begin{cases} (1-\alpha)\cdot\omega(x,y)_{n,i} + \alpha & \text{for } (11) \\ (1-\alpha)\cdot\omega(x,y)_{n,i} & \text{otherwise} \end{cases} \tag{12}$$

where:
$\omega(x,y)_{n+1,i}$ – weight approximated for i-th distribution,
$\omega(x,y)_{n,i}$ – previously approximated weight,
α – estimation factor, usually $\alpha \in (0.01, 0.1)$.

The analyzed area is considered foreground if the condition (11) is not satisfied for any distribution. In such case, the distribution with the lowest weight is replaced by a new distribution with the mean value equal to the mean value of the area in question, large standard deviation and low weight [8].

This method also works quickly without requiring a huge amount of memory, but its biggest advantage is the ability to model a varying background through the use of several Gaussian distributions. Its main limitation is the inability to correctly estimate the background which variation excesses the predetermined number of distributions.

The third way to estimate the histogram faces this restriction. The algorithm can adjust the number of Gaussian distributions in use to the modeling conditions [9]. This method is different from the previous one in only three places [9]:

- algorithm starts with a single Gaussian distribution,
- if none of the currently used distributions satisfies the update condition, and their number is less than the maximum, then a new distribution is added; if the maximum number of distributions is reached, a new distribution replaces the one with the lowest weight,
- if the weight of one of the distributions in use falls below the minimum threshold, the distribution is removed.

In this approach defining whether the analyzed area belongs to the background is based only on the distributions with the highest weights.

The method obtains better results than the previous one often in less time with a similar memory usage. This is due to the active change of the number of Gaussian distributions in use. Using this approach, it possible to model a highly varying background, as well as reduce the amount of calculation, when the background is static.

The sample results of this algorithm are shown in figure 5. In comparison with the previously described approaches, this method is undoubtedly the best. The spot noise appears only if the light conditions change rapidly. The moving objects are very well reproduced, and there are no trails or shadows remaining behind them.

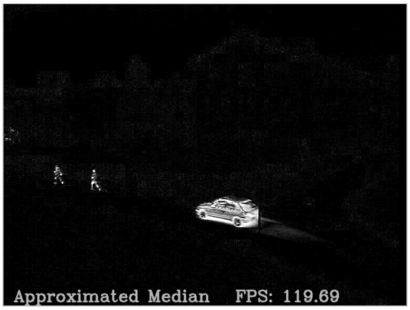


Fig. 4. Results returned by the approximated median method

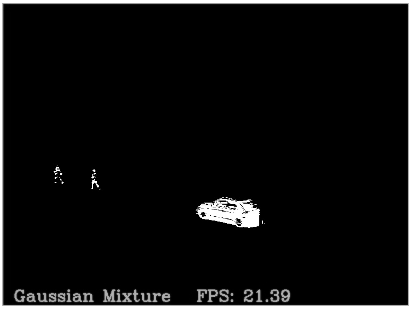


Fig. 5. Results returned by varying Gaussian distributions number method

2.6 The Modeling Using the PCA Algorithm

A completely different approach is presented in the algorithm described in [10]. It uses the Principal Component Analysis as a tool for the background modeling. In this case, the PCA analyzes the principal components of the background image.

The PCA algorithm can reduce the dimensionality of the data set, that is, in the context of the background modeling, the quantity of the background features. This is achieved by transforming the coordinate system to maximize the variance of the subsequent coordinates. In consequence, several initial data dimensions carry most of the information of the entire data set. This allows omitting the dimensions with the lowest variance. In the context of the image processing the coordinates are the subsequent pixels; a single data sample is the entire image, and a data set is a database of images. Based on eigenspace created using the PCA it is possible to recreate every input data sample while the new samples can be recreated only partially or not at all.

In the described method the database, for which the principal component analysis is conducted, is created from the initial M frames of the analyzed video sequence. A detailed description of the PCA algorithm flow is presented in [10]. Here I will only note that the moving objects do not occur in the same locations in the frames used to create the eigenspace, so that they do not have a significant impact on the model. Thanks to this feature, the image areas containing the moving objects cannot be restored on the basis of the model, while the static areas are described by the model as a linear combination of eigenvectors. This means that the eigenspace is a very good model describing exclusively the areas that belong to the background [10].

According to the authors, this method returns good results in less time than the Gaussian mixture methods described earlier. Due to the computational complexity, this method lies between the approximating methods and the methods that analyze N previous frames. The algorithm major limitation is the inability to adapt the model to the changing appearance of the background. In order to do so, it would be necessary to conduct entire PCA from the beginning, which is computationally very expensive.

3 Conclusion

3.1 The Properties of the Described Modeling Methods

The three most important features for the evaluation and comparison of the described algorithms are:

- Processing speed
  The algorithms approximating the mean and the median proved to be the fastest. Other methods are several times slower, but are still able to work in real time.
- Amount of memory required
  The algorithms approximating the mean and the median are also the best in this aspect. They need to remember only two frames at a time. The methods using the Gaussian distributions and the PCA use from about a dozen up to several dozen frames at once. The most resource consuming are the statistical approaches analyzing N previous frames of the sequence.
- Accuracy of results returned
  There is no objective measure of the accuracy of the results of the background modeling so far. The authors of the outlined algorithms evaluate the results of their work subjectively by comparing them to the results of other approaches, which generally comes down to determining the amount of noise and distortions caused by shadows and ghosts objects. Such criteria are best meet by the methods using the Gaussian distributions and the PCA. The increased amount of noise and distortions are present in approaches approximating the mean and the median as well as in those which calculate the statistical values based on N previous frames of the analyzed sequence; the accuracy of the results are similar in all of them. The least accurate are methods detecting the movement areas based on a simple difference of successive frames of the analyzed sequence.

3.2 Further Development Possibilities

Almost all of the described methods, as well as the vast majority of all other approaches, either ignore or do not take into account several important issues:

- The information from the subsequent chromatic channels is considered separately, while by definition all color components explicitly define it, so they should be considered together. Among the outlined approaches only the vector median considers all of the chromatic channels together.
- The information about the location of a pixel in the image is overlooked together with its correlation with its neighborhood, which is undoubtedly present, for example, in the form of a color gradient. The only approach that takes into account the spatial correlations present in the image is the method using the PCA.

- There is a certain freedom left in the case of the background model update. This raises the question of when, if at all, to consider a pixel located for a long time in the same place and belonging to the foreground to be part of the background. With this sub-problem, the phenomenon of the ghost objects, or groups of pixels not belonging to the background model, which should belong to it, is combined. The ghost objects form as a result of removing a part of the scene which has been classified as the background, for example a car driving away from a parking lot lefts an empty parking space, which could become a ghost object.
- The problem of the shadows cast by the objects present in the scene is also considered rather lightly. The shadows are a phenomenon that makes it difficult to classify areas as the background. The shaded background areas, although clearly different from the non-shaded ones, should also be considered as the background.

Due to the current state of the most commonly used, best described and researched methods, the new background modeling algorithms, in addition to having the advantages of the described approaches, should also take into consideration (and appropriately use) the color information as well as the spatial correlations in the image and properly update the background model in order to compensate for the shadows and not to allow the formation of excessive noise and ghost objects.

References

1. Picardi, M.: Background subtraction techniques: a review. The ARC Centre of Excellence for Autonomous Systems, Faculty of Engineering, University of Technology, Sydney (2004)
2. McIvor, A.M.: Background Subtraction Techniques. In: IVCNZ 2000, Hamilton, New Zealand (2000)
3. Parks, D.H., Fels, S.S.: Evaluation of Background Subtraction Algorithms with Post-processing. Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada (2008)
4. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting Objects, Shadows and Ghosts in Video Streams by Exploiting Color and Motion Information. In: Proc. of the 11th International Conference on Image Analysis and Processing, ICIAP 2001, pp. 360–365 (2001)
5. Zhang, L., Liang, Y.: Motion human detection based on background subtraction. In: Second International Workshop on Education Technology and Computer Science. IEEE (2010)
6. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Statistic and knowledge-based moving object detection in traffic scenes. In: Proc. of ITSC 2000, pp. 27–32 (2000)
7. Arsić, D., Hristov, E., Lehment, N., Hörnler, B., Schuller, B., Rigoll, G.: Applying multi layer homography for multi camera person tracking. In: Proc. of AMMCSS 2008, Stanford, CA, USA (2008)
8. KaewTraKulPong, P., Bowden, R.: An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection. In: Proc. of 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS 2001 (2001)

9. Zivkovic, Z.: Improved adaptive Gausian mixture model for background subtraction. In: International Conference Pattern Recognition, UK (August 2004)
10. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. IEEE PAMI 22, 831–843 (2000)
11. Yasira Beevi, C.P., Natarajan, S.: An efficient Video Segmentation Algorithm with Real time Adaptive Threshold Technique. International Journal of Signal Processing, Image Processing and Pattern Recognition 2(4) (2009)
12. PETS2001 dataset, The University of Reading, UK (2001)

A Hybrid Fusion Technique for Watermarking Digital Images*

Ahmed S. Salama[1], Mohamed A. Al-Qodah[1], Abdullah M. Iliyasu[1],
Awad Kh. Al-Asmari[1], and Fei Yan[2]

[1] Salman Bin Abdulaziz University,
P.O. Box 173, Al Kharj 11942, Kingdom of Saudi Arabia
[2] Department of Computational Intelligence & Systems Science,
Tokyo Institute of Technology, Japan

Abstract. A hybrid fusion technique (HFT) that offers improved imperceptibility, better robustness to attacks and superior quality in terms of watermark image recovery than is realisable using the traditional Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT) digital watermarking techniques or methods that utilise their direct combination is presented in this work. The separate DWT and DCT methods embed the watermark image directly on the wavelet or cosine coefficients of the cover image, while methods that combine both the DWT and DCT are certain to directly embed the watermark on some parts of the cover image twice. Unlike all these methods, our proposed HFT technique spreads the watermark in the transform (wavelet and cosine) domains separately and then the watermarked images from the two images emanating from the two domains are fused together to generate the final watermarked image - producing a hybrid. Such hybridisation allows us to exploit the individual benefits derivable from the separate use of the DWT and DCT methods. Experimental evaluations show that combining the two transforms in this manner offers improved performance in terms of imperceptibility, robustness to jpeg and other kinds of attack on the published (watermarked) image and quality of recovered watermark image than is obtainable using the methods that are based solely on the DWT or DCT techniques or their direct combination. The proposed HFT technique guarantees additional protection for digital images from illicit copying and unauthorised tampering.

Keywords: Digital image, watermarking, data hiding, information security, discrete wavelet transforms, discrete cosine transform, image fusion, hybrid fusion technique.

1 Introduction

With the increasing use of internet and effortless copying coupled with the ease of tempering and unauthorised distribution of digital data, the need for copyright

* This work is supported by funding from the Salman Bin Abdulaziz University, Saudi Arabia Special Research Grant (SAU/UCI-EE001.1434).

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_20, © Springer International Publishing Switzerland 2014

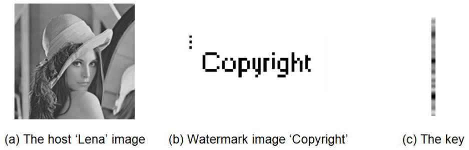
enforcement technologies that can protect copyright ownership of multimedia objects is both paramount and exigent. Digital watermarking is widely accepted as a technique for labelling multi-media data, including digital images, text documents, video and audio clips, by hiding secret information in the data [1, 2, 3].

There are a lot of digital image watermarking techniques all of which can be broadly categorised into two major classes: spatial-domain and frequency-domain watermarking techniques [4]. In the spatial domain techniques, the values of the image pixels are directly modified on the watermark which is to be embedded. The least significant bits (LSB) technique [5], one of the earliest such techniques, is implemented by modifying the least significant bits (LSB) of the image's pixel data. While in frequency-domain techniques, the transform coefficients are modified instead of directly changing the pixel values. To detect the presence of a watermark, the inverse transform is used. Commonly used frequency-domain transforms include the Discrete Wavelet Transform (DWT), the Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT). However, due to its excellent spatial localisation and multi-resolution characteristics, which are similar to the theoretical models of the human visual system, DWT [6] has been more frequently used in digital image watermarking. Further performance improvements in DWT-based digital image watermarking algorithms could be obtained by combining DWT with DCT [7]. This method, which we will henceforth refer to as the DWT+DCT method, relies on a sequential combination of the wavelet and cosine transforms which transform the images, resulting in some parts of the host image being transformed twice. First of all the entire image is decomposed into its DWT sub-bands and then depending on the algorithm used, some of these sub-bands are further transformed using the DCT method. The idea stems from the fact that combined transforms could compensate for the drawbacks of each other, resulting in more a effective watermarking technique. Exploiting this objective our proposed technique offers improved imperceptibility, a key benefit of DWT methods, while also providing better robustness to attacks, which is a core advantage derivable from DCT watermarking techniques. Unlike the other methods that sequentially combine DWT and DCT methods [7, 3], our proposed technique fuses the watermarked images obtainable from the watermarking the same host image separately using the DWT and DCT methods. In addition to improving the imperceptibility and robustness (to attacks on the watermarked images), the proposed technique offers better recovery of the watermark (from the watermarked images) than is possible using the standard (separate DWT or DCT) methods and those that directly combine the DWT and DCT methods.

The outline of the remainder of the this paper is as follows: the details of the proposed technique to improve the imperceptibility and robustness of the watermarked images including the watermark embedding and extraction procedures required to accomplish it are presented in Sect. 2. This is followed, in Sect. 3, by the discussion and analysis of experimental results obtained using the proposed technique.

2 The Proposed HFT Technique

The proposed two-tier, hybrid fusion (HFT) technique including its watermark embedding and extraction processes are presented in this section. Our proposed technique fuses the watermarked images realised from the separate application of the DWT and DCT methods on the same host image. In the first step, the DWT method, the content of the host image is decomposed using the Discrete Wavelet Transform (DWT) with Haar filter. To make the resulting image more robust against attacks, the watermark signal is then inserted into two regions of the third level sub-band of the cover image [8]. The widely used 'Lena' test image will be used as our cover (or host) image. While a simple 20×50 binary image will be used as our watermark signal (or image). In DWT methods, it is common practice to embed the watermark using one key in two regions in the detailed wavelet coefficients of the host image [8]. This is useful in order to improve robustness against several kinds of attack while also preserving the imperceptibility of the watermarked image. Occasionally, we shall refer to the trio of the host image, the watermark image and the key as simply I, W, and K respectively. These images are presented in Fig. 1. The proposed procedures for watermark embedding and extraction are summarised and presented in Fig. 2 and Fig. 3, respectively.



(a) The host 'Lena' image (b) Watermark image 'Copyright' (c) The key

Fig. 1. (a) the host Lena image, (b) the 'Copyright' watermark image and (c) the key

2.1 Watermark Embedding Algorithm

The watermark embedding process consists of three simple steps as enumerated in the sequel.

1. Apply DWT on the cover image as described in [8]
2. Apply DCT on the same cover image as described in [9]
3. Fuse the resulting images from 1 and 2 above to generate the watermarked image.

The fusion in step 3 above involves a pixel-wise collation of the image contents from the separate DWT and DCT watermarked versions of the same image.

2.2 Watermark Reconstruction Algorithm

To recover the watermark image from an already watermarked image, four simple steps, as enumerated below, are required.

1. Apply DWT on the watermarked image as described in [8]
2. Apply DCT on the same watermarked image as described in [9]
3. Correlate (pixel-wise) between the recovered watermark from DWT and the recovered watermark from DCT
4. Generate the recovered watermark image

The standard correlation operation is employed in a pixel-wise fashion to generate the recovered watermark image. The resulting watermarked Lena images presented in Fig. 4, and specifically the peak-signal-to-noise-ratio (PSNR) values, confirm that the quality of the watermarked Lena image obtainable using the proposed hybrid fusion technique (result in the far right) is better than those realisable from both the separate (DWT and DCT methods alone) and the combined (sequential combination of the DWT and DCT methods) method. Similarly, the proposed hybrid fusion technique offers better recovery of the watermark than is obtainable from both the separate (DWT and DCT methods alone) and the combined (sequential combination of the DWT + DCT) method. The recovered 'Copyright' text watermark image for the separate DWT, DCT, the combined (DWT + DCT) method and the proposed hybrid fusion (HFT) techniques are presented in Fig. 5.

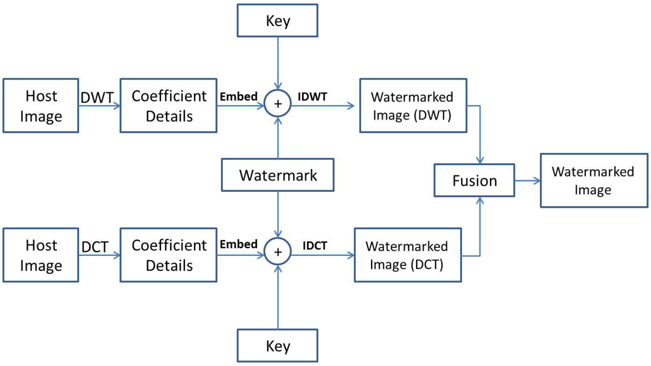


Fig. 2. Watermark embedding procedure for the proposed hybrid fusion (HFT) technique

3 Experimental Results and Analysis

3.1 Performance Study

In analysing the performance of image-hiding techniques, many parameters have been proposed. These parameters include visual quality (imperceptibility), complexity, payload capacity, execution time, robustness and a few others depending on the objective of the watermarking strategy [10]. From among them, we shall

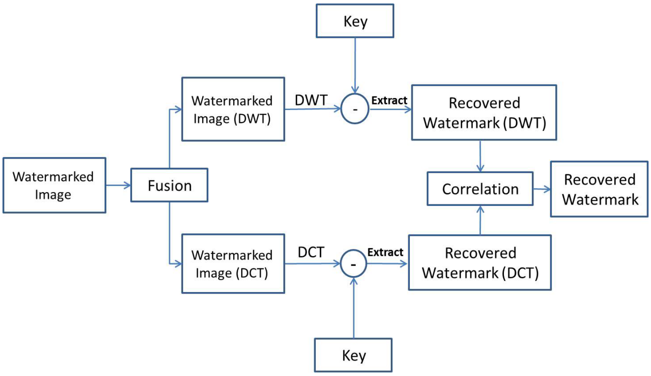


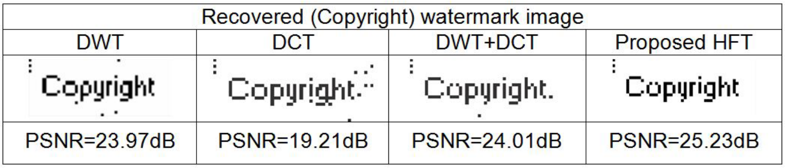
Fig. 3. Watermark extraction procedure for the proposed hybrid fusion (HFT) technique

limit the criteria adopted for measuring the performance of our scheme alongside other image hiding methods to the trio of watermarked image quality (a.k.a. imperceptibility), the execution time (in terms of computing resources) and, finally, the ability of the watermarked image to withstand attacks, i.e. robustness. Therefore, the rest of this section is devoted to analysing the performance of our proposed hybrid fusion technique alongside the separate DWT and DCT methods and the combined (DWT and DCT) method using the aforementioned criteria.

Original Lena image	Watermarked Lena image			
	DWT	DCT	DWT+DCT	Proposed HFT
	PSNR=49.29dB	PSNR=29.58dB	PSNR=39.27dB	PSNR=37.01dB

Fig. 4. The watermarked Lena image as realised from the separate DWT, DCT, the combined (DWT + DCT) method and the proposed hybrid fusion techniques

Measuring Imperceptibility (Perceptual Quality). The core measure here is the numerical PSNR values obtained using each of the four methods under analysis (and also the visual quality of the watermarked images themselves). But due to brevity the inherent distortions in the watermarked versions obtained using each of the four methods are not easily visible, hence, we constrain the comparison to the numerical PSNR values). Using the results presented in Fig. 5, it is clear that DWT gives the best result. This is attributed to the fact that the method of embedding watermarks, such as in CDMA [8], involves inserting the watermark in the third level of DWT image sub-band (i.e. in the



Recovered (Copyright) watermark image			
DWT	DCT	DWT+DCT	Proposed HFT
Copyright	Copyright	Copyright	Copyright
PSNR=23.97dB	PSNR=19.21dB	PSNR=24.01dB	PSNR=25.23dB

Fig. 5. The text 'Copyright' watermark images as recovered from the separate DWT, DCT, the combined (DWT + DCT) method and the proposed hybrid fusion techniques

LL2 of HH sub-band). It has been proven that embedding the watermark in this region does not degrade the quality of the watermarked image. In other words, the imperceptibility of the watermarked image is unaffected. Since this region contains non-significant bits inserting the watermark does not erode the quality of the watermarked image, i.e. its imperceptibility. Nonetheless, the proposed hybrid fusion technique, in second position, fares better than the other three methods under analysis. Interestingly, after varying the dimensions of the image and watermark signal we saw that the PSNR values obtained using the proposed hybrid fusion technique is approximately the average of the values obtained using the separate DWT and DCT methods. Unsurprisingly, since it has long been established that the DWT method gives better image quality than DCT [2], further experiments here found that comparison-based correlation in DCT mid-band has less PSNR values than the two previous methods.

Measuring Execution Time. This experiment is designed to determine the computational cost of each watermarking algorithm. A desktop computer with i5 2. 67 GHz CPU, 4 GB Ram equipped with the necessary softwares was used as our simulation environment. For each algorithm, CPU timings are extracted for the watemark insertion and extraction procedures for each image. The results of this test are summarised in Fig. 6. As seen therefrom, the DCT technique requires the least computational resources for its execution, while the DWT technique, in which the watermark is embedded two times in two regions (HL3 and LH3 in LL2 of the HH), after which the correlation between the extracted watermarks is done, takes the second longest time. As expected, the proposed hybrid fusion technique proved marginally slower than the separate DWT and DCT methods. It takes approximately the sum of the time taken the separate DWT and DCT methods. In terms of the extraction procedure this slight lag in time is attributed to the correlation step as explained in Sect. 2 (specifically Fig. 3). Remarkably though, the proposed hybrid fusion technique is marginally faster than the method that sequentially combines the DWT and DCT methods. These results are summarised in the chart in Fig. 6.

Measuring Robustness. In order to determine how well each algorithm can survive distortions, the robustness of the implemented algorithms against signal processing operations are tested using three different types of attacks: the JPEG compression, blurring, and median filtering. These are chosen because they don't

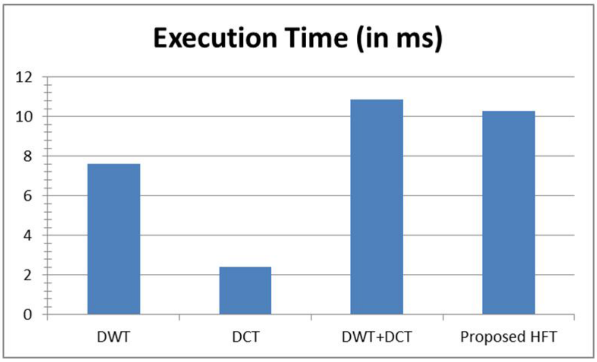


Fig. 6. Comparison of the execution time for the DWT (DWT), DCT (DCT), the combined DWT and DCT (DWT+DCT) and proposed HFT techniques

severely degrade the subjective quality of the watermarked images and because they represent real world types of operations.

Some of the early literature considered a binary robustness metric that only allows for two different states; the watermark is either robust or not. However, it makes sense to use a metric that allows for different levels of robustness. The use of the bit-correct ratio (BCR) has recently become common, as it allows for a more detailed scale of values. The bit correct ratio (BCR) is defined as the ratio of correct extracted bits to the total number of embedded bits and can be expresses using the formula:

$$BCR = \frac{100}{l} \sum_{n=0}^{l-1} \begin{cases} 1, & W_n^{'} = W_n \\ 0, & W_n^{'} \neq W_n \end{cases}$$

where l is the watermark length, W_n corresponds to the n^{th} bit of the embedded watermark and $W_n^{'}$ corresponds to the n^{th} bit of the recovered watermark.

- JPEG compression

For each algorithm, the embedding procedure is applied on each host image; the watermarked images are then compressed using different percentages of quality ranging from 10 to 90 percent. A plot of the bit correct ratio (BCR) as a function of percentage quality is shown in Fig. 8 below. Therefrom, we see that the proposed hybrid fusion technique (HFT) fares better than the separate DWT (DWT) and the combined DWT and DCT (DWT+DCT) methods but fares slightly worse than the DCT method. This suggests that the proposed hybrid fusion technique is slightly fragile to JPEG compression while the DCT method outperforms all the other methods due to the fact that it is the core component of the JPEG compression.

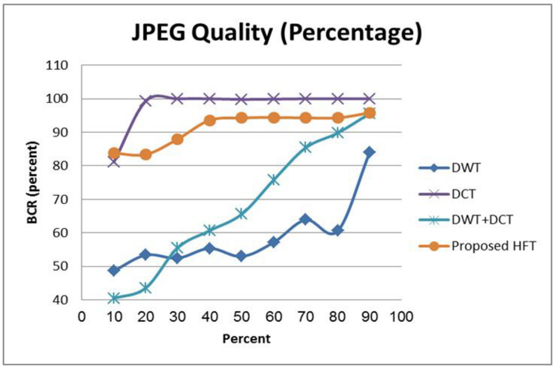


Fig. 7. Comparison for the Bit correct ratio due to JPEG compression

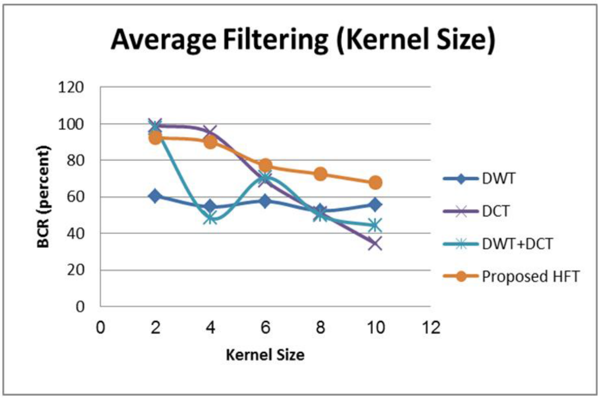


Fig. 8. Bit correct ratio due to blurring of the watermarked images

– BCR for blurring

For blurring, a mean filter of size K×K samples is applied to the watermarked images, for 2≤K≤10, to replace each pixel with the average of a block from the surrounding pixels. The results of this experiment are shown in Fig. 8. As seen therefrom, the BCR for the proposed hybrid fusion technique is close to that of the DCT for kernel sizes varying from 2 to 6. The BCR for the DWT method is somewhat linear while the combined DWT and DCT (DWT+DCT) manifests a slightly unstable variation with increase in the kernel size.

– BCR for median filter

Median filtering is a non-linear process used to reduce high frequency noise in an image. A median filter of size K×K, for 2≤K≤10, is used to replace each sample of the watermarked image with the median value from the set of neighbouring

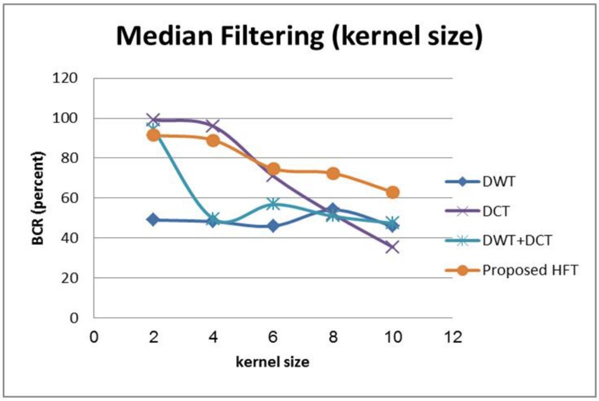
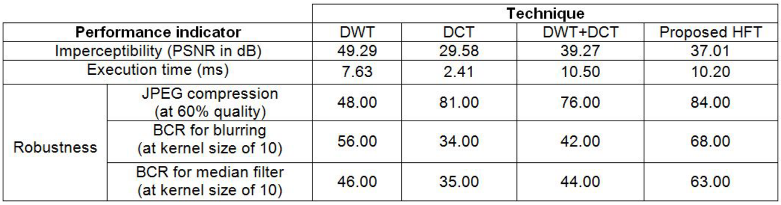


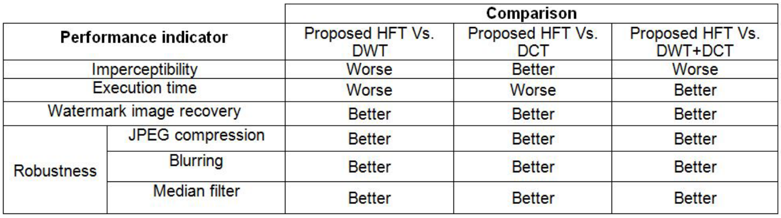
Fig. 9. Comparison for the bit correct ratio from median filtering of watermarked images for the four methods - DWT, DCT, DWT+DCT and HFT methods



		Technique			
Performance indicator		DWT	DCT	DWT+DCT	Proposed HFT
Imperceptibility (PSNR in dB)		49.29	29.58	39.27	37.01
Execution time (ms)		7.63	2.41	10.50	10.20
Robustness	JPEG compression (at 60% quality)	48.00	81.00	76.00	84.00
	BCR for blurring (at kernel size of 10)	56.00	34.00	42.00	68.00
	BCR for median filter (at kernel size of 10)	46.00	35.00	44.00	63.00

Fig. 10. Performance evaluation based on imperceptibility, execution time and robustness for the four methods under review: DWT, DCT and DWT+DCT alongside the proposed hybrid fusion (HFT) technique

pixels. The results comparing the four methods under analysis based on these settings are shown in Fig. 9. This figure suggests that the proposed HFT method is the most robust against median filtering. In median filtering, the neighbouring pixels are ranked according to brightness (intensity) and the median value becomes the new value for the central pixel. It does well with suppressing impulse noise, in which some individual pixels have extreme values. Using our proposed HFT, the complete removal of these individual pixels will not significantly affect the hidden watermark because it is hidden in both DCT and DWT coefficients. The table in Fig. 10 presents a numerical summary of the overall performance evaluation based on the three performance metrics; imperceptibility, execution time and robustness for the four methods under review, i.e. the DWT, DCT and DWT+DCT, alongside our proposed hybrid fusion technique. It is apparent therefrom that the proposed technique has the best robustness (compression, blurring and filtering) than the other three methods. Although it fared worse than the DCT method in terms of its imperceptibility it still maintained PSNR values that are approximately the average of those using DWT and DCT methods and is still better than the DWT and DWT+DCT methods. The trend is similar in terms of computational time required to execute the four methods. Remarkably, the proposed HFT method outperforms all the other methods in terms of faithful reproduction of the watermark image. The comparison between the proposed HFT technique and the other three methods as discussed earlier in this section are succinctly summarised in the table in Fig. 11 below.



		Comparison		
Performance indicator		Proposed HFT Vs. DWT	Proposed HFT Vs. DCT	Proposed HFT Vs. DWT+DCT
Imperceptibility		Worse	Better	Worse
Execution time		Worse	Worse	Better
Watermark image recovery		Better	Better	Better
Robustness	JPEG compression	Better	Better	Better
	Blurring	Better	Better	Better
	Median filter	Better	Better	Better

Fig. 11. Summary of the performance of the proposed HFT technique alongside the other three methods based on imperceptibility, execution time, watermark image recovery and robustness

4 Conclusions

Frequency-domain watermarking methods such as the discrete wavelet transform (DWT) and the discrete cosine transform (DCT) have been applied successfully in many digital image watermarking techniques. Separately, the DWT and DCT methods have been proven to improve the fidelity of watermarked images (a.k.a imperceptibility) and the ability of the watermarked image to resist attacks, commonly referred to as robustness, respectively. Methods that exploit these individual qualities to realise improvements in the watermark image quality indices have been suggested. However, these methods just combine the two transforms sequentially with certain parts of the image transformed twice. In this paper, we have described a hybrid - DWT and DCT - technique that fuses the watermarked images realised by the separate watermarking of the same image using the DWT and DCT methods. The proposed HFT technique offered improved performance in terms of imperceptibility, robustness to jpeg and other kinds of attack on the published (watermarked) image and quality of recovered watermark image than is obtainable using the methods that are based solely on the DWT or DCT techniques or their direct combination. An interesting perspective to improve the performance of the proposed watermarking technique is to accord extra attention to the perceptual visual complexity inherent to the image being watermarked [11, 12].

References

[1] Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography. Elsevier (2008)

[2] Langelaar, G., Setyawan, I., Lagendijk, R.: Watermarking digital image and video data: A state-of-art overview. IEEE Signal Proc. Mag. 17(5), 20–46 (2000)

[3] Al-Haj, A.: Combined dwt-dct digital image watermarking. Journal of Computer Science 3(9), 740–746 (2007)

[4] Potdar, V., Han, S., Chang, E.: A survey of digital image watermarking techniques, Perth, Australia. In: Proc. of the IEEE Int. Conf. on Industrial Informatics, pp. 709–716 (2005)

[5] Neeta, D., Snehal, K., Jacobs, D.: Implementation of LSB steganography and its evaluation for various bits. In: 1st Int. Conf. on Digital Info. Management, pp. 173–178 (December 6, 2006)

[6] Vetterli, M., Kovacevic, J.: Wavelets and Subband Coding. Prentice Hall, USA (1995)

[7] Rao, K., Yip, P.: Discrete Cosine Transform: algorithms, advantages, applications. Academic Press, USA (1990)

[8] Salama, A., Atta, R., Rizk, R., Wanes, F.: A robust digital image watermarking technique based on wavelet transform. In: IEEE Int. Conf. on Sys. Eng. and Tech., pp. 100–104 (2011)

[9] Langelaar, G.C., Setyawan, I., Lagendijk, R.L.: Watermarking digital image and video data. IEEE Signal Processing Magazine 17, 20–43 (2000)

[10] Iliyasu, A., Le, P., Dong, F., Hirota, K.: Watermarking and authentication of quantum images based on restricted geometric transformations. Information Sciences 186(1), 126–149 (2012)

[11] Le, P., Iliyasu, A., Garcia, J., Dong, F., Hirota, K.: Representing visual complexity of images using a 3D feature space based on structure, noise, and diversity. J. of Adv. Computational Intelligence and Intelligent Informatics 16, 631–640 (2012)

[12] Al-Asmari, A., Salama, A., Iliyasu, A., Al-Qodah, M.: A DWT ordering scheme for hiding data in images using pixel value difference. In: IEEE Eighth Int. Conf. on Computational Intelligence and Security (CIS), pp. 553–557 (2012)

Metadata Projection for Visual Resources Retrieval

Jolanta Mizera-Pietraszko

Institute of Informatics,
Wroclaw University of Technology
jolanta.mizera-pietraszko@pwr.wroc.pl

Abstract. The amount of multimedia documents available in the net stimulates a strong need to explore metadata for improving efficiency of systems' performance whether commercial, prototypical or open source. This paper provides a framework on image retrieval in cross-language environment based on metadata. Digital images provide visual resources information content that is known as content-based image retrieval (CBIR). We investigate the correlation between query linguistic characteristics that match the visual object collections and the image databases as well as some indexing strategies. Our test set forms a baseline to analyze influence of English and French phenomena on the strategy of browsing the Web for image collections in multilingual environment. Metadata in our experiment is defined as keywords, annotations, image captions or descriptors.

Keywords: Image processing, information retrieval, information technology.

1 Introduction

Digital image or visual object collections are widely used in medicine, university studies, business, media and many more domains. Apart from the images these collections usually contain other visual resources like graphics, paintings, photos or drawings.

From the user viewpoint the most common techniques are browsing a collection one by one, searching it by specifying a specific image characteristics like size, colour, resolution which is known as a direct image querying, or eventually querying a collection by example called as image pattern. Pragmatically, quite common is semantic technique which relies on browsing a digital collection for an object specified by a user [10].

In the process of multi-language information retrieval, in which the system retrieves the images annotated in all languages occurred in the databases crawled by an engine, and cross-language IR (such a system translates the keywords in order to search for visual resources annotated in a language indicated by the user), the key strategies focus on both translation quality of the user's query as well as the adequacy of the digital content to the user's need accordingly.

These two factors influence the overall system performance to almost the same degree. Any mismatch at the stage of the source language translation is a serious

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_21, © Springer International Publishing Switzerland 2014

impediment to the number of visual documents retrieved. Simultaneously, any ambiguity occurring in the query affects the system performance as well. Fuzzy idea about the user's need is the other decisive factor for inadequate system responses.

To propose a new approach to image collection indexing that results in achieving almost human-like quality translation we decided to analyze one of the most popular tool Alta Vista Babel Fish service based on the Systran system, that is known as the first MT system that produces the highest translation quality.

At first, we present a brief introduction to machine translation models in general, section two outlines the cross-lingual search engine characteristics essential for the experiment, then we proceed with introduction of our linguistic test set structure based on which the images are clustered into categories accordingly. We also provide example-based evaluation methodology of translation quality since to each category added is an example structure being a model which is then used to train our system.

In the next section we discuss our distortion-sensitive model used subsequently for our analysis of the relation between the subcategories' co-occurrences in the corpora and their automatic translation quality.

Finally, we consider metadata-oriented image retrieval incorporating maximum likelihood method for these two language pairs. We discuss the impact of the metadata for our main linguistic categories to find on the visual retrieval once for English and then for the same French queries to come to the conclusion of our metadata image retrieval bijective analysis.

2 AltaVista Multilingual Capacities

The name of AltaVista derives from "a view from above" in 1995 by a Digital Equipment Corporation's Research Lab in Palo Alto, California, US. Soon it became a first multilingual search engine which deployed language recognition, the biggest index for browsing multimedia digital collections.

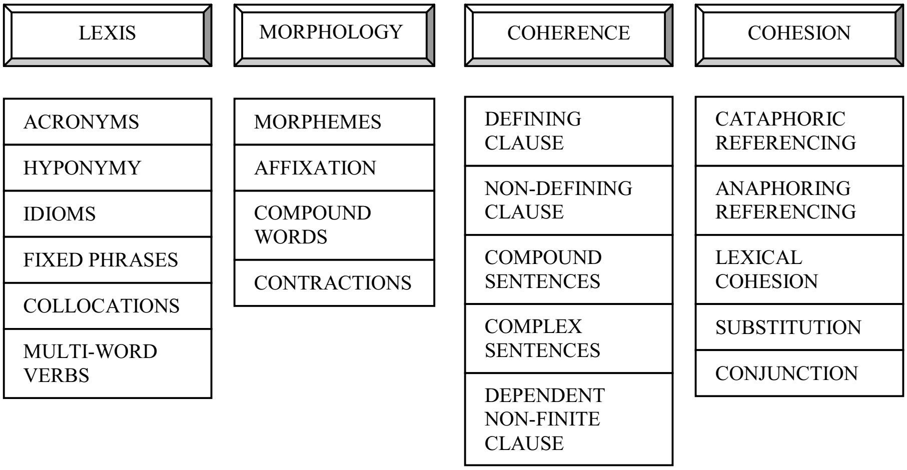
Its component Babel Fish provides translation services as the first one on the web and actually it facilitates translation of phrases and the whole web pages of all the European Union language pairs [11].

The search technology relies on analysis of digital document characteristics like a content of a web page, its title, description, source and the links to the query. For visual documents these characteristics are quite often a bag of words with no special meaning.

In multimedia databases relevance is heavily dependent upon interpretation of the visual content which can vary significantly for most of the users [13]. Likewise, the metadata like captions, description, annotations or the document title somewhat limit the range of interpretations.

3 The Test Set Structure

As Figure 1 shows, we constructed a test set that consists of grammar structures with example queries as multimedia annotations in order to investigate relevant to the user's need identification of multimedia objects and finally its retrieval process.



LEXIS	MORPHOLOGY	COHERENCE	COHESION
ACRONYMS	MORPHEMES	DEFINING CLAUSE	CATAPHORIC REFERENCING
HYPONYMY	AFFIXATION	NON-DEFINING CLAUSE	ANAPHORING REFERENCING
IDIOMS	COMPOUND WORDS	COMPOUND SENTENCES	LEXICAL COHESION
FIXED PHRASES	CONTRACTIONS	COMPLEX SENTENCES	SUBSTITUTION
COLLOCATIONS		DEPENDENT NON-FINITE CLAUSE	CONJUNCTION
MULTI-WORD VERBS			

Fig. 1. Grammar structures of the test set

The structure of our test set covers four main grammar areas: lexical devices, morphology, coherence and cohesive devices [4] extracted from the corpora. The critical point at this stage of our experiment was to adopt such a methodology in each case that will result in making the subgroups absolutely separate from each other. Thus, we established the approaches to avoid any disambiguation.

In order to clarify our approach we define some of the categories [9]:

- acronym – initial letters of words or word parts in a phrase or a name
- morpheme – the smallest linguistic unit that has semantic meaning
- hyponymy – inclusion of a word subgroup
- fixed phrase – a group of words that function as a single unit in the syntax
- cataphoric referencing – co-reference of an expression with another one
- anaphoric referencing – use of articles to point backwards

4 Translation Results with Babel Fish Services

Our test set was submitted for translation to the Babel Fish services. Here are some example sentences of the translation process with the score in brackets:

(1) IDIOM He worked himself to the bone. - il s'est travaillé à l'os. - It was worked with the bone. (0.164)

(2) FIXED-PHRASE I never understood the ins and outs of his job.- Je n'ai jamais compris les coins et recoins de son travail. - I never understood the corners and recesses if its work. (0.337)

(3) MULTI-WORD VERB They didn't have anywhere to put us up so we stayed in a hotel. - Ils n'ont pas eu n'importe où pour nous mettre vers le haut ainsi nous sommes restés dans un hôtel. - They did not have to put to us anywhere to the top thus we remained in a hotel. (0.427)

(4) COMPOUND kind-hearted – bienfaisant - beneficial (0.418)

Here the first sentences are the examples of the categories labeled. They are the reference English into French translations whereas the last sentences are their hypothetical French into English translations.

The example sentences presented above show how much the translation process is irreversible. In the next step, each category was evaluated with Meteor, the metric that matches words called unigrams of hypothetical to its reference translation accordingly [3]. This metric has been recognized as more precise than Blue [2] and perhaps the closest of all to the human judgment. It has been used during the TREC and NIST campaigns over many years as the most precise metric for evaluation of translation quality.

4.1 Systran Translation Services

Systran, a producer of BabelFish translation services, uses the Interlingua model in their distributions. The engine translates texts of no more than 150 words long in 19 languages including French. In 2001, the producer introduced declarative programming that allows the designer to describe some language phenomena by graphical formalism rather than coding the task steps [8]. Interlingua model has been developed to implicit transfer based on parallel source and target descriptions. In addition, XML exchange format component was added. In 2004, a commercial system was tested on Spanish into English translation and scored 0.56 with Blue.

Before submitting our test set, we compared some of the translation results produced by Systran Professional 3.3 [5] to BabelFish services. Surprisingly, more than 90% of the French sentences was translated into English in the exactly the same way which indicates that the technique of processing the text has not been changed.

Some problems like different use of separators by the engine, document structure that is not preserved by the MT systems, or poor parser performance cause the same errors in both systems compared [6]. Obviously, the commercial Systran distributions are provided with a range of facilities not included in the on-line translation services.

4.2 Lexis Measured with Meteor

Our test set consisted of examples (texts, word lists, or sentences), each relating to a different feature, scored 0.7105 with the Porter stem module and 0.6935 with the Exact module. However, using the newest version of Meteor, our test set scored 0.6663 with the Exact module only. The difference is a result of replacing constant parameters with the variable ones in order to make the evaluation closer to the human judgment.

Due to the fact that Meteor reports only the best score out of all the segments assessed, we divided the test set into subsets of sentences belonging only to one category of the discourse analyzed. The aim was to find the features that are translated with the highest accuracy and those with the lowest one.

Table 1. Evaluation of the translation quality of the lexical features

LEXIS	Acronyms	Hyponymy	Idioms	Fixed phrases	Collocation	Multiword verbs
Score	0.315	0.463	0.164	0.337	0.490	0.427
Precision:	0.333	0.444	0.300	0.333	0.500	0.476
Recall:	0.375	0.500	0.333	0.363	0.555	0.588

The table shows a difference between evaluation of the text as a whole and its crucial points in particular. We observe that the features which occurs the most often achieve relatively better results than the others. Therefore, the highest score goes to the morphemes and the lowest one to the acronyms and idioms.

4.3 Morphology Measured with Meteor

This section presents distribution of morphological features. To be consisted with the ESOL project, we decided to include morphemes despite they are, in fact, a part of affixation so that to make them a separate feature, we agreed to exclude the suffixes from it.

Table 2. Evaluation of translation quality of the morphological features

MORPHOLOGY	Morphemes	Affixation	Compound words	Contractions
Score	0.5282	0.4259	0.4189	0.1639
Precision:	0.5000	0.5000	0.4286	0.1429
Recall:	0.5714	0.5000	05000	0.1667

Despite a lack of contractions which have been removed from the text, a few extracted from the whole corpora were translated by the Babel Fish services reasonably well like e.g. couldn't, or don't.

4.4 Coherence Measured with Meteor

On the contrary to cohesion, coherence addresses semantic meaning of the sentences.

Table 3. Evaluation of translation quality of the coherence features

COHERENCE	Defining Clause	Compound Sentences	Complex Sentences	Main Clause	Simple Sentences	Dependent Nonfinite Clause	Nondependent Clause
Score	0.415	0.810	0.614	0.747	0.225	0.371	0.606
Precision	0.588	0.782	0.571	0.702	0.416	0.416	0.562
Recall	0.555	0.818	0.666	0.764	0.454	0.416	0.642

Both complex sentences as well as dependent non-finite clauses achieved the highest score. However, almost all the features were processed correctly by the system.

4.5 Cohesion Measured with Meteor

Cohesive devises correlate grammatical aspects with lexis of the text. This study does not deal with ellipses as the rarest cohesive devise used by the speakers and too difficult to be recognized by the text analyzers.

Table 4. Evaluation of translation quality of the cohesive devices

COHESION	Cataphoric Referencing	Anaphoric Referencing	Lexical Cohesion	Substitution	Conjunction
Score	0.7397	0.5324	0.543	0.3156	0.685
Precision:	0.6923	0.6111	0.500	0.4000	0.695
Recall:	0.7500	0.6111	0.625	0.4000	0.695

Cataphoric referencing is not that common as the anaphoric one as it is a way of introducing a subject in an abstract way and then addressing it directly. For substitution we extracted words like "one/ones". Lexical cohesion was based on definite articles and determiners. Grammatical conjunction scored high as it is correctly translated by the most MT systems.

5 Distortion-Sensitive Translation Model

In this section we analyze reliability of the linguistic units in correlation to their co-occurrences in the Speakers' turns. In other words, the objective is to determine the extent to which frequency of the grammatical structures analyzed corresponds to their translations. Reliability of image retrieval $\mathcal{R}\{\mathfrak{T}(u_i),\mathfrak{F}(u_i)\}$ is defined here as an absolute value of a distance function of the co-occurrences $\mathfrak{T}(u_i)$ of a particular linguistic unit u_i in the corpora to the approximated by METEOR value of the translation accuracy $\mathfrak{F}(u_i)$.

$$\mathcal{R}\{\mathfrak{T}(u_i),\mathfrak{F}(u_i)\} = \wedge_{i=1}^{n}|\mathfrak{T}(u_i) - \mathfrak{F}(u_i)|$$

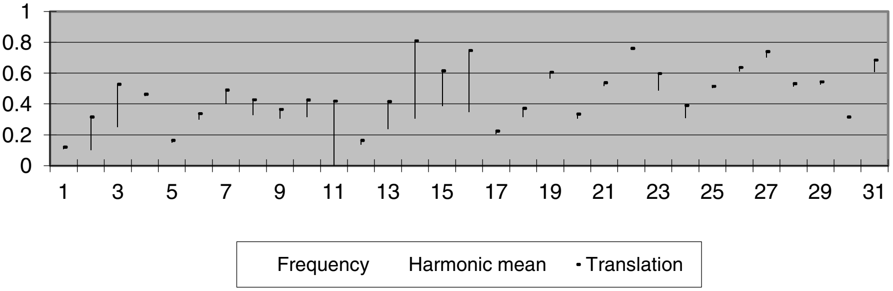


Fig. 2. Distance-based model of co-occurrences of the linguistic units and the translation accuracy

Consequently, the diagram above presents the distance governed by the mathematical formula described that is defined as an absolute value of frequency of the grammatical units and their translation quality measured with Meteor.

In the middle point of the distance the harmonic mean is a weighted average of precision and recall calculated by Meteor and known as 1-Factor. The horizontal axis shows the features' numbers 1 to 31 while the vertical axis, their values taken from the tables above. The length of the vertical lines indicates the distance.

As seen, in most cases there is a point instead of a line which suggests equality of these two values which resembles a close correspondence between the unit co-occurrence and its translation accuracy. Therefore, we observe that for the greatest number of our linguistic units the occurrence is simply proportional to the translation quality.

6 Metadata-Oriented Image Retrieval

Following a definition from Encyclopedia of Database Systems [1] "multimedia metadata is structured, encoded data that describes content and representation characteristics of information-bearing multimedia entities to facilitate the automatic or semiautomatic identification, discovery, assessment, and management of the described entities, as well as their generation, manipulation, and distribution".

We focus on the entities which represent linguistic units and therefore project retrieval efficiency in English and French. Multimedia metadata represent image captions, keywords, annotations as well as entity descriptions. From the users viewpoint, it is a challenge to describe multimedia content in their queries with the aim to improve matching process carried out by AltaVista engine. However, it may be a comparison basis for other engines with cross-language search functionality.

In the next step, all the translation results were submitted to AltaVista search engine for image retrieval. Each category of the English phrases was submitted separately to a set of French images and then vice versa. Relevant documents were only those that included the whole query in the metadata. Since the research purpose is to evaluate the correlation between translation quality and the number of the relevant visual search results retrieved from numerous databases, for each query we calculate precision (*Pr*) as in Table 5 and Table 6.

For *N* units $x_1,..,x_n$ where θ denotes the estimated probability *Pr,* our joint density function satisfies the equation

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times f(x_3 \mid \theta) \times \ldots \times f(x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

By the means of Maximum likelihood *(ML)* we compute the mean and the standard deviation (*SD*) of the normal distribution [12] against the number of our search results ranging from 0 to 1,200,000 for each language pair separately.

Figures 3 and 4 show the histograms of the number of our results and the curve of the fitted normal distribution with its maximum point at our Mean value $M(\overline{x_i})$=9188.5357 for EN-FR and $M(\overline{x_i})$=10947.2000 for FR-EN image retrieval where

$$SD(\overline{x_i}) = {}^{SD(\theta)}\!\Big/_{\sqrt{N}} \qquad \text{while} \qquad SD(SD_i) = {}^{SD(\overline{x_i})}\!\Big/_{\sqrt{2N}}$$

Table 5. Statistical Parameters for EN-FR Maximum Likelihood Approximation

Parameter	Estimated θ	Standard Deviation
Mean $M(\overline{x_i})$	$M(\overline{x_i}) =$ 9188.5357	$SD(\overline{x_i})$=7329.4989
Standard Deviation	$SD(\theta)$ =38784.0626	*SD(SD$_i$)*=5182.7383

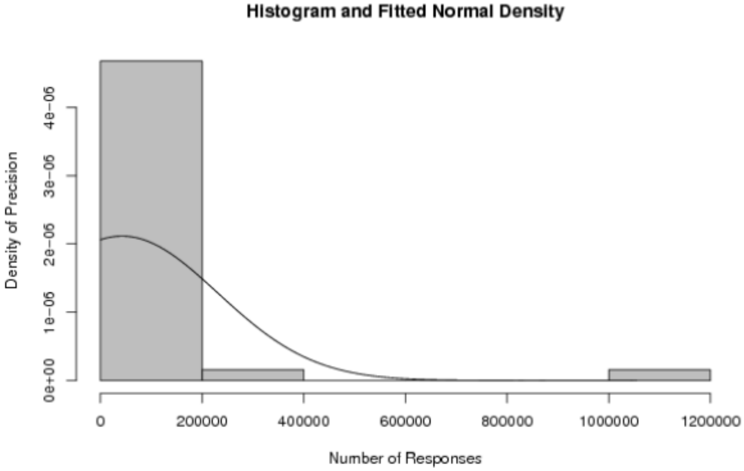


Fig. 3. A histogram of the visual search results for EN-FR translations [14]

Table 6. Statistical Parameters for FR-EN Maximum Likelihood Approximation

Parameter	Estimated θ	Standard Deviation
Mean $M(\overline{x_i})$	$M(\overline{x_i})$ =10947.2000	$SD(\overline{x_i})$=10704.2232
Standard Deviation	$SD(\theta)$ =58629.4451	*SD(SD$_i$)*=7569.0288

These two graphs look quite similar. All the queries with Lexis analyzed achieve definitely the best translations and consequently, the system scores the highest accuracy and precision. The second best appears to be cohesion features (the last histogram bar in these two figures), but the values are quite small when compared to the lexical features.

Morphology in English databases submitted by a set of French queries proved to perform much better than from the French visual collections (see Figure 4, bar two). The curves drop dramatically for all so called longer queries that are those consisting of more than five query words depending on their co-occurrences in the databases searched. Observed is a great discrepancy between the number of the system visual

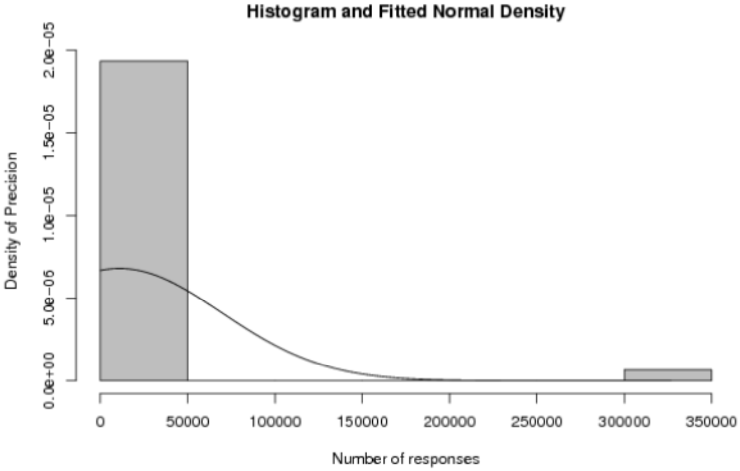


Fig. 4. A histogram of the visual search results for FR-EN translations [14]

responses, from a million to zero for many linguistic features (compare the parameters' values for these two language pairs). The relevance is definitely not at all dependent upon the number of responses – for some relatively small system responses it rises up to 50% whereas for those with around a million ones it is estimated to 10% only.

However, the databases include far more images with English rather than French visual documents (see the figure axes above). At this point the ratio of the French number images is approximated up to thirty percent of the English visual documents.

7 Conclusion

Metadata image retrieval is incomparable to text-based document searching. Thus, not only the system technology or the translation model deployed, but it is the user behavior, especially his or her interaction with the system that determines the set of relevant responses. For more experienced or proficient users it is far easier to operate a system in a way that results with the image ranking irrespective of their e.g. annotations or any metadata.

Another finding of our experiment indicates that not all of the linguistic structures analyzed here impede the query translation process e.g. coherence devices, adjective-adverb rules in particular used as metadata for image search. Although, it is an essential issue to relate the factors that influence the limitation of the system responses we should concentrate only on those that cover the area of the user's interest.

In our further study we plan to concentrate on some other translation models in order to analyze the performance of the system particular components while processing the linguistic features specified for visual objects annotated in a couple of languages deriving from the same language group.

References

[1] Nack, F.: Multimedia Metadata. In: Liu, L., Ozsu, T. (eds.) Encyclopedia of Database Systems. Springer Science+Business Media (2009)

[2] Callison-Burch, C.: Re-evaluating the role of BLUE in MT Research. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, CACL (2006)

[3] Baneriee, S., Lavie, A.: Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics, Linguistics, Michigan (2005)

[4] Halliday, M., Rugayia, H.: Cohesion in English. Longman (1976)

[5] Lewis, D.: PC-Based MT; An Illustration of Capabilities in Response to Submitted Test Sentence. MT Review No. 12, The Periodical of the Natural language Translation, Specialist Group of the British Computer Society, Issue No. 12 (2001)

[6] Senellart, P.: Systran Translation Stylesheet (2000)

[7] Lioma, C., Ounis, I.: Deploying part-of-speech patterns to enhance statistical phrase-based machine translation resources. In: ACL 2005: Workshop on Building and Using Parallel Texts – Data-driven Machine Translation and Beyond, pp. 163–166. University of Michigan, Ann Arbor (2005)

[8] Senellart, J., Boitet, C., Romary, L.: Systran New Generation. In: MT Summit IX, pp. 22–26 (2003)

[9] Swan, M.: Practical English Usage. Oxford University Press (2000)

[10] Town, C., Sinclair, D.: Language-based querying of image collections on the basis of an extensible ontology. Image and Vision Computing 22(3), 251–267 (2004)

[11] AltaVista at a Glance – Yahoo, Europe Ltd. London, UK

[12] Grefenstette, G. (ed.): Cross-Language Information Retrieval. Kluwer Academic Publisher Boston, US (1998)

[13] Crestani, F., Lalmas, M., van Rijsbergen, C.: Information Retrieval: Uncertainty and Logics. In: Advanced Models for the Representation and Retrieval of Information. Kluwer Academic Publishers (1999)

[14] Wessa, P.: Free Statistics Software, Office for Research Development and Education (2009)

Automated Detection Type Body and Shape Deformation for Robotic Welding Line

Pavol Božek

Slovak University of Technology,
Faculty of Materials Science and Technology,
Trnava, Slovakia
pavol.bozek@stuba.sk

Abstract. Car production starts at a mill where the steel plates are cutted out to body parts. This is followed by a fully automated welding shop, where robots weld the different body parts. All models are welded on the same line. At the end of the line is automatic control of dimensions. In line samples are welded chassis all models, robots are still "see" how the car looks like. Then every car body is still checked in detail and continues to paint shop. After painting the body continues to assembly hall where about 35 % of the work carried out by robots and other activities make installers: Replacing the interior and exterior parts, engine mounting, axle, exhaust system and other mechanical components.

Keywords: automation, defect, detection, production system.

1 Introduction

Production logistics is part of each sector. At present the automotive industry resonates in Slovakia at the first place. It is therefore necessary to deal with the quality, type, or shape deformation of each distributed piece of bodywork at the end of the production line. We compare the pair of car bodies where one is the etalon and the other is tested and compared with a reference standard model in order to detect errors or body classified to the appropriate category. Finally, add wheels, seats and steering wheel [12].

During the production phase, it is necessary to monitor and control the production process and detect possible errors incurred, if necessary sort by different body models and distribute them according to the type where it is needed. For these purposes, it is appropriate to use automated system that can compare the model with just the right product manufactured products and detect any inaccuracies (eg. by welding the various parts of the body, where it is compared with sample pieces produced). Another possible application is an automated system for distribution into body assembly hall – breakdown by type of bodywork.

Experiments with a fully automated system of recognizing objects met with varying levels of success. But none of them reached such a high level of accuracy that can be run completely unattended.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_22, © Springer International Publishing Switzerland 2014

Automatic recognition of objects dedicated work of many authors. The authors of [10] present the results of the integration of the proposed system for automatic object recognition, based on the decomposition of objects. Describe the possibility of recognizing objects independently of their position, orientation and size. The authors [14] discuss the issues and technologies for automated compilation of object models and sensors to optically recognizing strategies for detecting and determining the position of the object. The authors [15] proposed the use of advanced computational intelligence and distributed processing to mimic the human brain's thinking in recognizing objects. If it is combined with a cognitive process of detection technologies, the system can combine traditional techniques with image processing computer intelligence to determine the identity of different objects. The authors [16] proposed a stereo vision system, which should provide information on the design project efficiently, quickly and inexpensively. Describe a framework for rapid 3D object recognition technology.

The contribution describes the design of the verification methods for the correct classification of the car bodies using Fourier-Mellin transformation. Consequently, the images are compared using Fourier transformation and phase correlation.

For comparison, or to determine the degree of similarity or images we used the variety of metrics, for example phase correlation and the percentage comparison.

2 Mathematical Principle

Fourier-Mellin transformation allows comparison of images which are offset, rotated and have changed scale. This method takes advantage of the fact that the shift differences are annulated because amplitude spectrum of the image and its displaced copy are identical, only their phase spectrum is different. Subsequent log-polar transformation causes that the rotation and scale will appear outwardly as a shift, so that phase correlation can be used to determine the angle of rotation and scaling between a pair of images during their registration. Phase correlation is based on the fact that two similar images in cross-spectrum create continuous sharp extreme in the place of registration and the noise is distributed randomly in discontinuous extremes [1,7].

2.1 Registration of Images

It is the reference image $a(x)$ and the input image $b(x)$, which has to be the identical with the reference image. The registration function of geometric transformation is to be estimated from the similarity of the characteristics of these images.

Consider that the image of $b(x)$ is the displaced copy of image $a(x)$:

$$b(x) = a(x - x_0), \tag{1}$$

their Fourier's transformation $A(u)$ and $B(u)$ have a relationship:

$$B(u) = e^{-2\pi(u^t x_0)} A(u). \tag{2}$$

We can construct a correlation function [6]:

$$Q_p(u) = \frac{A^*(u)}{|A(u)|} \cdot \frac{B(u)}{|B(u)|} = e^{j(\theta_b(u)-\theta_a(u))} \tag{3}$$

where $\theta_a(u)$ and $\theta_b(u)$ are phases of $A(u)$ and $B(u)$.

In the absence of noise, this function can be expressed in the form:

$$Q(u) = e^{-j2\pi(u^t x_0)}. \tag{4}$$

Its inverse Fourier transform is Dirac δ-function centered in the [6]:

$$u = m_0 = [x_9, y_0]^t. \tag{5}$$

Registration is accomplished by detecting the occurrence of Dirac δ-function in the inverse transformation of function $Q_p(u)$. The coordinates of the maximum culmination of δ determine the image translation.

In practice the noise in the image $Q_p(u)$ can be complicated by the search for global maximum [3]. Therefore, it is advantageous to use the low pass filter, the weight function which "mutes" high frequencies (noise). The result is a matrix that has a clear peak, whose position (deviation from the center) corresponds to the mutual displacement of the images.

The disadvantage of this method is that, without further adjustment it is not possible to register other transformation than the shifts. When registering holographic images it is necessary to synchronize the images not only to each other but also images rotated, or with modified scale.

Modification of the above method – use of the Fourier-Mellin transformation – allows registration of shifted or rotated images and with different scaling.

Fourier-Mellin transformation combines aspects of the Fourier and Mellin transformation with the transformation into a log-polar coordinates of the image.

Registration of images using Fourier-Mellin transformation uses phase and amplitude. This method uses the fact that the differences of shifts are ignored, because the amplitude spectrum of the image and its displaced copy is identical, only their phase spectrum varies.

The rotation can be converted to the shift transformation of images into a polar coordinate system. However, we need to know the center of rotation, which, in practice, of course, is unknown. This problem can be eliminated by working with the amplitude spectrums of images.

If the picture b is rotated on an angle with respect to the image a, the amplitude spectrum $|F(b)|$ against the spectrum $|F(a)|$ is rotated about the same angle. However, in this case the center of rotation is known – is it the point representing the zero frequency.

If the amplitude spectrums $|F(a)|$, $|F(b)|$ are transformed to the log-polar coordinate system (the spectrum is converted to polar coordinates and the distance from the origin of the coordinate system to the logarithmic scale), by the above described method of the phase correlation we identify not only the rotation, but also change of the scale.

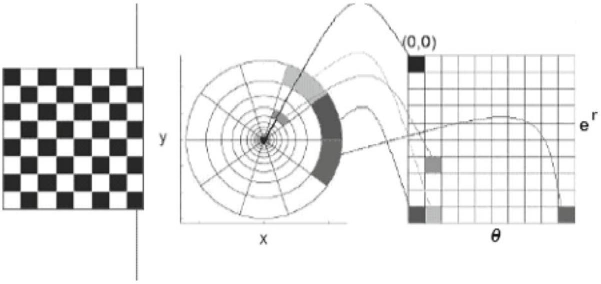


Fig. 1. Log-polar transformation [4]

Fourier-Mellin transformation converts the rotation and zooming to easy shifts in the parametric space and allows the use of the techniques of the phase correlation. Phase correlation then can be used to determine the angle of rotation and scale between the pair of images.

Picture function $f(x, y)$ may be sampled as a function $f(\theta, e^r) = f(\theta, \rho)$, where r is the distance from the center of the image see Fig. 1.

Suppose that the centre of the image is the starting point for the transformation. Each pixel in the image can be represented as the distance r from the center of the image and the angle θ. If we rotate the image, only θ is changed, r remains the same.

If instead of a representation of the second pixel coordinate as the amount of r we use the exponential scale log r, we can convert the change of scale to translation.

If the image has been resized to scale according k, the Cartesian point $P(x, y)$ in the image will be in log-polar coordinates represented as $P(\theta, \log(k \cdot r))$. Then the point P with the changed scale will be expressed as translation: $P(\theta, \log k + \log r)$ [8].

Used conversion from Cartesian to the log-polar coordinates:

a) Log-polar transformation of the amplitudes $|A(u, v)|$, $|B(u, v)|$ from Cartesian to the log-polar coordinate system Fourier transformation is displayed in log-polar plane by the transformation of coordinates see: Fig. 2.

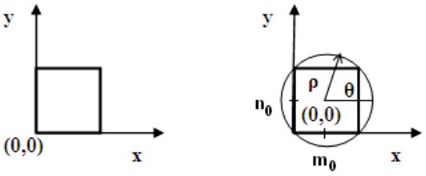


Fig. 2. Transformation of rectangular coordinates to polar by [2]

Origin (m_0, n_0) should be in the middle of the image matrix, to ensure the maximum number of pixels. If the image is formed by a square grid of $N \times N$ points, the coordinates of the center will be:

$$\begin{array}{lll} m_0 = N/2; & n_0 = N/2 & \text{if } N \text{ is odd} \\ m_0 = (N-1)/2; & n_0 = (N-1)/2 & \text{if } N \text{ is even} \end{array}$$

Maximum sampling radius for conversion will be:

$$\begin{array}{l} \rho_{\max} = \min(m_0, n_0) \ldots \text{ inscribed circle,} \\ \rho_{\max} = \sqrt{(m_0^2, n_0^2)} \ldots \ \ \text{described circle} \end{array}$$

If the inscribed circle is selected as the limit for conversion, some pixels, which lie outside of the circle will be ignored. If it described circle is selected, all the pixels will be included, but also defective pixels will be included (pixels inside the circle, but outside of the picture matrix). Whereas the pixels in Cartesian coordinates cannot be mapped one to one to the log-polar coordinates, the average of surrounding pixels (nearest neighbor, bilinear or bicubic downsampling) must be calculated.

Relationship between polar coordinates (ρ, θ), which is sampling the input image to the log-polar image (e^r, θ) is given by:

$$(\rho, \theta) = (e^r, \theta).$$

For pixel mapping from the input image (x_i, y_i) to pixels of the output image (ρ_m, θ_n) applies [2]:

$$\begin{array}{l} x_i = \text{round}(\rho \cdot \cos(\theta_n) + m_0), \\ y_i = \text{round}(\rho \cdot \sin(\theta_n) + n_0), \end{array} \tag{6}$$

where $(\rho_m, \theta_n) = (e^{rm}, \theta)$ by (8). The input image is of dimension $i \times j$ and the output image is of $m \times n$ dimension.

b) Fourier transformation of log-polar amplitudes

$$\begin{array}{l} A_{lp}(\nu, \varpi) = \mathcal{F}\{|A_{lp}(e^r, \theta)|\}, \\ B_{lp}(\nu, \varpi) = \mathcal{F}\{|B_{lp}(e^r, \theta)|\}, \end{array} \tag{7}$$

Log-polar transformation of amplitude spectrum causes the rotation and scaling to arise as the shift. It is therefore possible to use the phase correlation to detect the angle of rotation and scale between the pair of images.

Using phase correlation of the results of the Fourier-Mellin transformation A_{lp}, B_{lp}, we find the rotation size and scale of the test image b against a reference image a. By backward rotation and scaling the test image b we create image b'. Then we calculate the Fourier transformation of the image b' and the reference image a. Using phase correlation we calculate displacement of images. Backward shift of the image b' creates image b''.

The designed algorithm of registration of images using Fourier-Mellin transformation

(1) Load of the Img1 (reference) and Img2 (input)

[(1a) Preprocessing of input images from (1)
Locate areas of interest in the image, and move to the center of the image]

[(1b) Hamming window for input images from (1 or 1a)]

(2) The calculation of the fast Fourier transform (FFT) for Img1 and Img2 from (1), (1a), (1b)

(2a) Extraction of amplitudes from (2)

(3) The transformation of the amplitudes from (2a) to the log-polar coordinates (using bicubic interpolation)

(4) Calculation of the FFT for amplitude from (3)

(4a) Amplitude extraction from (4)

(4b) Phase extraction from (4),

(5) Phase correlation for SCALING and ROTATION from the phases (4b) [Gaussian low pass filter]

(6) The detection of maxima dX, dY of the phase correlation from (5)

—– REGISTRATION —–

(7) The calculation of the scale ρ from the value of dY from (6)

(8) Calculate the angle of rotation θ from the value of dX from (6)

(9) Backward rotation and change the scale of the test image (using bicubic interpolation)

(9a) The change of rotation of Img2 from (1) of angle: – rotation from (8)

(9b) The change of the scale of the backward rotated Img2 from (9a) of 1/scale from (7)

(9c) Complement or trimming the size of modified Img2 from (9b) to the size of Img1 from (1)

[(10) Hamming window for registered Img1 from (1) and Img2 from (9c)]

(11) Calculation of the FFT for registered images from (10)

(11a) The Extraction of phases from (11)

(12) Phase correlation for the SHIFT of phases (11a) [Gaussian low pass filter]

(13) Calculation of displacement Δx, Δy as the deviation of the maxima of the phase correlation of (12) from the center of the correlation matrix

(14) Backward shift of Img2 from (9c) according to (13)

(15) Output of parameters of image transformations $(\Delta x, \Delta y, \rho, \theta)$

—– The end of REGISTRATION —–

Note: In square brackets [] there are optional parametric parts of the algorithm [11].

In Fig. 3 there is the amplitude and the phase spectrum of the Fourier transformation of the image of the car body and its Fourier's spectrum in log-polar coordinates.

2.2 Image Compariso

After registration of the images it was necessary to compare two images of bodies and determine whether they are the same model or not.

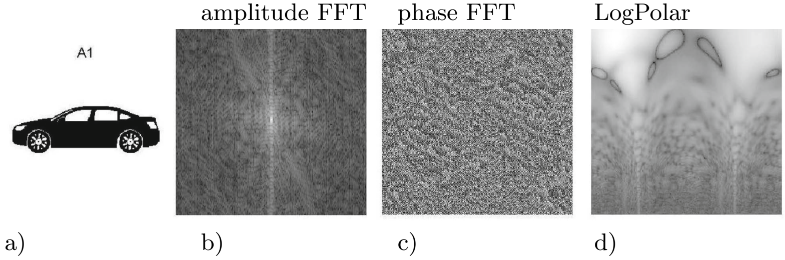


Fig. 3. a) Original image, b) Fourier spectrum amplitude in the Cartesian coordinates, c) Fourier spectrum phase in the Cartesian coordinates, d) Fourier spectrum in log-polar coordinates

We compared the reference designs with test images in order to determine the degree of similarity or correlation between them.

To evaluate the results of the comparison of the images we used several usual or custom metrics for the calculation of comparative score that quantifies the similarities between the test and the reference image. Calculation of metrics has been verified in different combinations of the application/without application of the hamming window, application/without application of the low-pass filter.

Some of the metrics used to determine similarity between reference and test image:

POC, Phase Only Correlation

Comparative score was calculated as the maximum phase correlation by [3]:

$$C = F^{-1}\left(\frac{F_a F_b^*}{|F_a||F_b|}\right),$$

where

- F – Fourier transformation of images a, b,
- F^{-1} – inverse Fourier transformation,
- F^* – complex conjugate image.

When the two images are similar, their phase correlation gives a distinct maximum. When the two images are not similar, it will create more insignificant maximum. The size of the maxima is used as a measure of similarity between two images.

MPOC, Modified Phase Only Correlation

Since the energy of the signal is lower in the high-frequency domain, phase components are not reliable in high-frequency domain. The effect of unreliable phase components in high frequencies can be limited by using filters or modifying of the POC function using spectral weighting function.

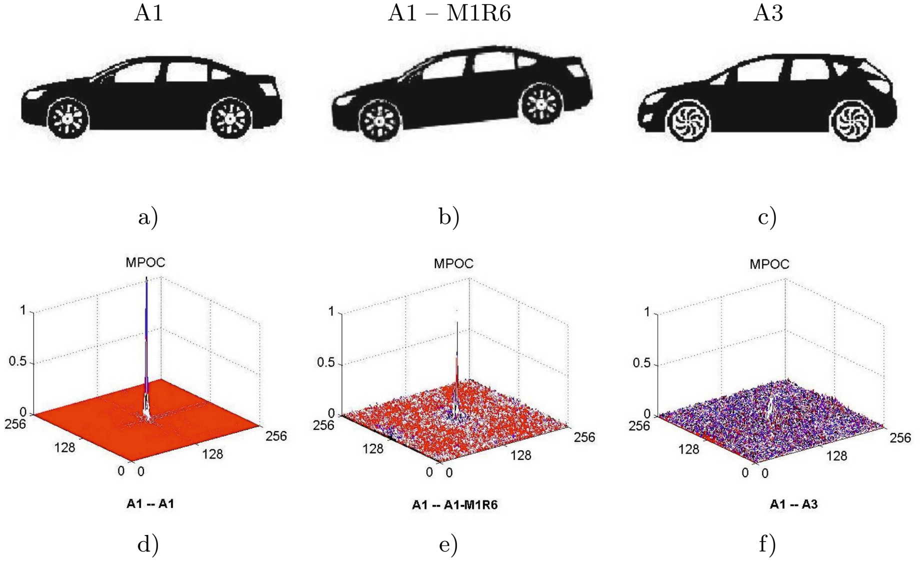


Fig. 4. Using of Modified Phase Only Correlation (MPOC): a) reference image, b-c) tested images, d) MPOC between identical images (a-a), e) MPOC between similar images (a-b), f) MPOC between various images (a-c)

To improve the detection by removal of minor ingredients with high frequency, which have a low reliability, the function of spectral weighting $W(u, v)$ has been used [5].

$$W(u,v) = \left(\frac{u^2+v^2}{\alpha}\right) e^{-\frac{u^2+v^2}{2\beta^2}},$$

where u, v are 2D coordinates, β is parameter, which checks width of function and α is used only to normalize.

Such modified image phase correlation function of a and b is given by [5]:

$$\tilde{q}_{a,b}(x,y) = F^{-1}\left\{W(u,v)\frac{F_a(u,v)F_b^*(u,v)}{|F_a(u,v)||F_b(u,v)|}\right\}.$$

The extreme value of a function $\tilde{q}_{a,b} = (x, y)$ is invariable at the change of shift and brightness.

This value was used to measure the similarity of images: if two images are similar, their function MPOC gives a crisp extreme, if they are different, then the extreme decreases considerably. Graphs were displayed in the range 1–N for coordinates x, y and functional values of MPOC normalized to the range 0–1 see Fig. 4.

PD, Percent Discrimination

Relative amount of similarity between reference and test image according to [6]:

$$PD = \frac{2[C_{ab}]_{\max}}{2[C_{aa}]_{\max} + 2[C_{bb}]_{\max}} \times 100\,\%$$

where $[C_{aa}]_{\max}$, $[C_{bb}]_{\max}$ and $[C_{ab}]_{\max}$ are maximal phase correlations if it is compared to a reference image $a(x, y)$ with itself, tested image $b(x, y)$ with itself and the reference image $a(x, y)$ with the image $b(x, y)$.

2.3 Decision

The calculated score is compared to the verification threshold, which will determine the degree of correlation necessary for comparison, which is to be taken as match.

On the basis of tests carried out, the threshold values of t have been set, according to which the system decision is regulated:

- images generating the result greater than or equal to t are evaluated as identical (this is the same car body),
- images generating results lower than t are evaluated as non-compliant (these are not the same car bodies).

The decisions of system can be: match, mismatch and without result, even though there are possible changing degrees of strong matches and mismatches.

The number of properly rejected and properly accepted images depends on the preset threshold value. The value is adjustable depending on the requirements, so that the system could be more or less accurate.

3 Functionality Verification of the Algorithm

The designs of car bodies from Shutterstock [9] have been used for functional testing of the algorithm. The images were adjusted to 256 shades of gray and to the dimensions of 256 × 256 pixels due to the use of the fast Fourier transform (FFT). The test images with the parameters of the transformation were created of these images. The images have been translated in the horizontal and/or vertical direction, rotated about the different angles in both directions and the scales have been changed. The images were compared with other car bodies.

The results of the assessment are given in Fig. 5. In the solved task it showed that the detection method used is applicable for the inclusion of the car bodies into the relevant categories.

The high success rate is achieved due to the used comparative set with rather small angle of rotation and scaling. Maximum correct recognition limit of this method has not been tested.

The images have not been properly recognized in the case of the scale too changed (0.7 and 1.3) in combination with translation and rotation.

The disadvantage of the proposed method can be time computational complexity. To register the image is to be calculated: 6× Fast Fourier Transform (FFT), 2× log-polar resampling, 2× phase correlation.

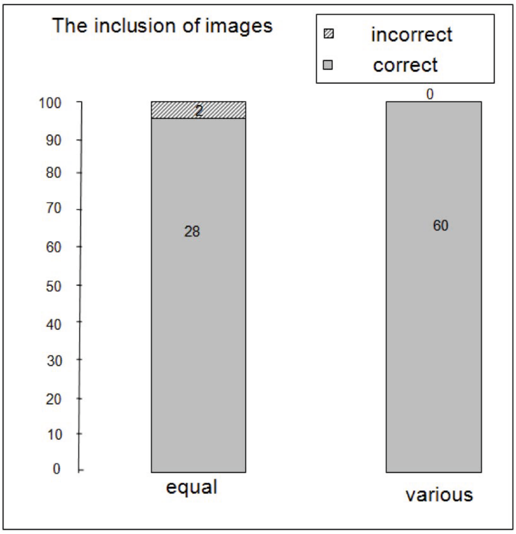


Fig. 5. Evaluation of the identification of images

4 Conclusion

Monitoring process provides information about the current state of construction, which can then be compared with the original model. Comparison is used to decide on change management in implementing the project. Current methods for retrieving and updating information on the progress of the project use digital cameras and laser-based systems.

The post describes the design of the verification method for the correct classification of the car body using a Fourier-Mellin transformation and subsequent comparison of the images using the Fourier transformation and phase correlation. Fourier-Mellin transformation offers image transformation resistant to the translation, rotation and scale. Method uses the fact that the integral transformations have their transformants in the case of translation, scaling and rotation, in the frequency area. In automatic processing it is possible to compare the images of the car bodies and find out if it is the same car body or not. It is possible to use different criteria for match, for example phase correlation, the difference correlation, the correlation coefficient, percentage comparison and comparison of calculated values with the chosen threshold for the relevant criterion. In our experiments we have set thresholds so that all the wrong couples of body works are revealed.

After adjustment algorithm can be used to compare the model with just the right product manufactured products in order to identify possible errors or omissions.

Acknowledgment. This work was written with a financial support VEGA agency in the frame of the project 1/0214/11 “The data mining usage in manufacturing systems control”.

This work was written as part of the strategic development program of the Izhevsk state technological university by name M. T. Kalashnikov PSR/A2/D2.5/ KAP "Development of methods for modeling and performance evaluation of manufacturing systems engineering as an information management system for automated machinery manufacturer".

References

1. Csongrády, T., Pivarčiová, E.: Spectral Biometrical Recognition of Fingerprints. Central European Journal of Computer Science 1(2), 194–204 (2011) ISSN: 1896-1533 (print version), ISSN: 2081-9935 (electronic version)
2. Dool, R.V.D.: Fourier-Mellin Transform. Image Processing Tools, http://www.scribd.com/doc/9480198/Tools-FourierMellin-Transform
3. Druckmüller, M., Antoš, M., Druckmüllerová, H.: Mathematical Methods for Visualization of the Solar Corona. Fine Mechanics and Optics – Science and Technology Magazine 50, 302–304 (2005)
4. Egli, A.: Medical Image Registration 2D/3D (X-Ray/CT), http://informatik.unibas.ch/lehre/fs09/cs503/_Downloads/egli.pdf
5. Gueham, M., Bouridane, A., Crookes, D.: Automatic Recognition of Partial Shoeprints Based on Phase-Only Correlation. In: IEEE, IV-441, ICIP 2007 (2007), http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4380049
6. Chen, Q.S.: Image Registration and its Applications in Medical Imaging. Dissertation work, Vrije Universiteit Brussel, Deconinck (1993)
7. Pivarčiová, E., Csongrády, T.: Fourier-Transform Mellinova – Tool for Image Registration. Aperiodikum Slovak Engineers and Scientists: Ideas and Facts XV(1-4), 18–22 (2009) ISBN 978-80-88806-79-0
8. Pratt, J.G.: Application of the Fourier-Mellin Transform to Translation, Rotation and Scale Invariant Plant Leaf Identification, Montreal (2000), digitool.library.mcgill.ca:1801/view/action/singleViewer.do?dvs=1244062994431~663&locale=sk&show_metadata=false&preferredextension=pdf&search_terms=000006662&adjacency=N&application=DIGITOOL-3&frameId=1&usePid1=true&usePid2=true
9. Shutterstock, Inc.: Stock Vector Illustration: Silhouette Cars on a White Background (2003-2012), http://www.shutterstock.com/pic.mhtml?id=111188486
10. Bennamoun, M., Boashash, B.: A Vision System for Automatic Object Recognition. In: IEEE International Conference on Systems, Man and Cybernetics: Humans, Information and Technology, vol. 2, pp. 1369–1374 (1994) ISBN: 0-7803-2129-4, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=400036&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D400036
11. Božek, P., Pivarčiová, E.: Registration of Holographic Images Based on Integral Transformation. Computing and Informatics 31(6), 1335–9150 (2012) ISSN 1335-9150
12. Bubeník, P.: Advanced Planning System in Small Business. In: Applied Computer Science: Managemen to Production Processes 7(2), 21–26, ISSN 1895-3735

13. Dvořák, F.: See how to Produce Cars kia in Slovakia. Auto, idnes.cz, http://auto.idnes.cz/podivejte-se-jak-se-na-slovensku-vyrabeji-vozy-kia-fs9-/automoto.aspx?c=A080916_151719_ automoto_fdv
14. Ikeuchi, K., Kanade, T.: Automatic Generation of Object Recognition Programs. Proceedings of the IEEE 76(8), 1016–1035, ISSN 0018-9219, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5972&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5972
15. Maddox, B.G., Swadley, C.L.: An Intelligent Systems Approach to Automated Object Recognition: A Preliminary Study. Open-File Report 02-461, USGS science for a changing world, http://egsc.usgs.gov/isb/pubs/ofrs/02-461/OFR02-461.pdf
16. Nuri, C., Hyojoo Son, S., Changwan, K., Changyoon, K., Hyoungkwan, K.: Rapid 3D Object Recognition for Automatic Project Progress Monitoring Using a Stereo Vision System. In: ISARC 2008: International Symposium on Automation and Robotic in Construction, pp. 58–63 (2008), http://www.iaarc.org/publications/fulltext/3_sec_010_Choi_et_al_Rapid.pdf

The Comparison of the Stochastic Algorithms for the Filter Parameters Calculation

Valeriy Rogoza[1] and Alexey Sergeev[2]

[1] Zachodniopomorski Uniwersytet Technologiczny, Szczecin, Poland
wrogoza@wi.zut.edu.pl
[2] NTUU KPI, ESC Institute of Applied System Analysis,
Kyiv, Ukraine
alexey.sergeev@ymail.com

Abstract. In this article the stochastic algorithms (particle swarm algorithm, simulated annealing algorithm, and genetic selection algorithm) applied to the problem of an adaptive calculation of the low pass filter parameters are compared. The data used for the filtration were obtained from the sensor (accelerometer) by implementing the software package for recording a human walking motion. For the algorithms comparison, the math library was implemented. The purpose of the study was to obtain optimum characteristics of moving average method by means of the algorithms described in this paper. The results of numerical experiments have shown that the best results have been obtained using the particle swarm algorithm and the genetic selection algorithm.

Keywords: filter, optimization, simulated annealing, particle swarm, genetic selection.

1 Introduction

The application of modern information technologies in the development of medical devices allows significantly to extend an applicability range of the latter and to improve an adequacy of patient's data analysis, reliability of conclusions about the state of his (her) health, and effectiveness of the proposed preventive and therapeutic methods [1]. Several diagnostic systems use sensors (accelerometers) for the registration of an acceleration of human's body parts. By means of accelerometers, extreme situations, such as falling of elderly persons [2], an excessive kinetic activity of persons with an unacceptable blood sugar level [3], drastic changes in body position, undesirable for people with a high blood pressure, etc., are also registered.

A more accurate estimation of the body acceleration is important in such problems, since it affects how the prediction of the human condition will be accurate and how the necessary actions will be taken promptly. Unfortunately, various magnetic, electrical and mechanical interferences have detrimental effects on the accuracy of data transmitted from the accelerometers to other devices in the diagnostic systems.

As a response to the mentioned interferences (noise), the sensor generates short-time electrical pulses of different amplitudes and durations. To depress these pulses, filtration is applied to raw accelerometer data. Since random pulses occur in the range

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_23, © Springer International Publishing Switzerland 2014

of high frequencies, it is appropriate to apply a low pass filter (LPF), which serves to separate the desired signal from noise.

Though, the processes registered by accelerometers are, as a rule, stationary, in case of moving with different speeds characteristics of these processes can considerably vary: for speed 0.926(m/s), movement frequency is 1Hz, acceleration amplitude is 3(m/s^2); for speed 1.032(m/s), frequency is 1.25Hz, amplitude is 4(m/s^2); for speed 1.316(m/s), frequency is 1.66Hz, amplitude is 6(m/s^2). In these circumstances a prior assignment of filter parameters is difficult.

To overcome these difficulties, one can suggest an approach of forming tunable filters. By tunable filter in this case, evolutionary synthesis of filter model based on the analysis of the signal from the accelerometer is meant. This approach assumes an adaptive filter tuning according to environmental conditions directly in the process of solving the filtration problem, i.e., the process of "learning" of the filter is combined with the process of solving the problem.

In this work, the comparative analysis of stochastic algorithms of the filter calculation which can form a basis of a tunable filter synthesis is performed, and practical recommendations for the use of these algorithms for the purpose of correcting filtration of accelerometer data are proposed.

2 Data Analysis

For the comparison of data filtering methods, human movements on a flat surface registered by the accelerometer were examined. The experiment task was to highlight the individual steps by analyzing the accelerometer signal incoming to processing device.

Since a movement of human body in space is uniform, the acceleration diagram is characterized by a certain cyclicity.

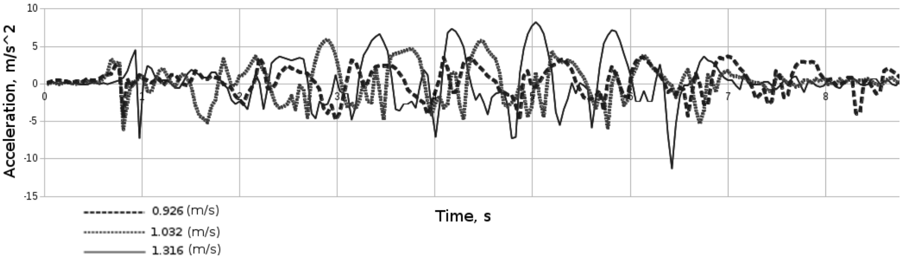


Fig. 1. Walking at a speed of 0.926 (m/s) = 3.312 (km/h) (7 control steps), 1.032 (m/s) = 3.708 (km/h) (6 control steps), 1.316 (m/s) = 4.716 (km/h) (5 control steps)

The signal amplitude is proportional to an acceleration which takes human body during walking (Figure 1), and the strong cyclical signal indicates a uniformity of human movement. An imposition of noise on the acceleration curve can lead to a significant distortion of the signal taken from the accelerometer, and thus significantly decrease the accuracy of the measured acceleration values. The noise component of the accelerometer signal is usually characterized by a higher frequency compared to the useful component, so a low pass filter (LPF) is included between the accelerometer and the processing device in order to eliminate the noise.

As noted above, the selection of the most suitable filter is not a trivial task, since human's locomotor functions and environmental conditions in which the experiments are performed, are quite different and unique in each case.

If acceleration values are determined at discrete time points, statistical methods of time series can be applied for the mathematical description of motion. An important characteristic of time series is autocorrelation coefficient:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2} \tag{1}$$

where rk is the coefficient of autocorrelation with the delay in the k measurements, $\overline{Y}$ is an average value of the time series, Y_t is observation at time t, Y_{t-k} is observation at time t-k.

In Figure 2, graphs of the time series autocorrelation coefficient are shown, defined for the three variants of human movement represented by graphs in Figure 1. It follows from the graphs of the auto-correlation coefficient (Fig. 2) that movement is a stationary process in all three variants.

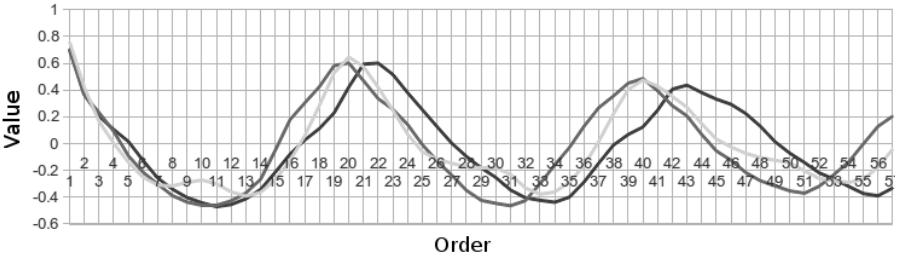


Fig. 2. The coefficients of autocorrelation for time series

Due to inertia, the accelerometer does not have time to track rapid changes of the instantaneous acceleration, so, even if there are other ideal conditions (absence of magnetic and electrical interference), locally noises occur, that can be mistaken for a real movement. As can be seen from Figure 1, these noises can be observed, for example, at the final points of the human movement. This signal, which can be considered as a noise of mechanical origin, can be filtered by the LPF as well as other types of noise.

When choosing a method of filtering, it is necessary to take into account the requirement for the preserving of correlation and regression which are present in the source data.

3 Filter Selection

The problem of synthesis of filters applicable to diagnostic systems of the human physical condition is discussed in a number of papers.

In [4], the filters (truncated mean, moving average, median) are compared on the data with different types of noise properties (uniform, exponential, and normal distribution).

The comparison of filtration results showed that the "truncated mean" filter is suitable in the case of uniform noise, the "moving average" filter is suitable in the case of noise with the normal distribution, and the median filter is suitable in the case of noise with an exponential distribution.

Since accelerometer data errors have the normal distribution, the "moving average" filter was selected for further data filtration.

4 Objective Function Selection

As a basis of the adaptive calculation of the moving average filter coefficients, the idea of calculation by using the method of least squares (LS) [5] was taken. The method of least squares was used to calculate the approximate data (as "ideal" data), which were compared with the smoothing data. The moving average filter was selected for the purpose of smoothing.

In general, the moving average filter is a special case of the weighted moving average filter, whose all the weights are equal to 1/n (n – number of measurements).

The equation for calculating the moving average filter (simple moving average filter) is as follows:

$$x'(t) = \frac{1}{n} \cdot \sum_{i=0}^{n-1} x(t-i) \tag{2}$$

where n is the number of previous time points taken into account, x(t) is the value of linear acceleration at moment t.

As can be seen from (2), the smoothed value is the average of the previous source data values. It depends on the choice of the value of n, and also on how much smoothing is sensitive to local fluctuations.

For small values of n, smoothing depends on the values of the last measurements and is sensitive to the large variations. For large values of n, local fluctuations are averaged, but filtering is done with a delay (by level).

To assess the relative quality of filtering, smoothed values were compared with the approximated values. In contrast to averaging, as in the case of "moving average", the regression (linear) relationship between the previous data is estimated in the process of approximation. Thus approximation allows us to save behavior of function and to reduce influence of the vibrational response. For optimal filtering, it is necessary to choose an appropriate value of n, i.e., to solve the optimization problem.

As the objective function to be minimized, the mean-square error value (MSE) was chosen. The relationship for the calculation of MSE was modified to compare the smoothed and the approximated values of acceleration.

The modified relationship of the mean-square error (MSE) for smoothed and approximated values is as follows:

$$MSE = \frac{1}{N} \cdot \sum_{t=1}^{N} (x'(t) - z(t))^2 \tag{3}$$

where N is the number of raw acceleration values, t is the time instance, x'(t) is the smoothed raw value, and z(t) is the approximated raw value.

The N smoothed and approximated source data values were chosen as the elements of the MSE equation given by (3). Approximation was made with LS and smoothing was made with the moving average filter.

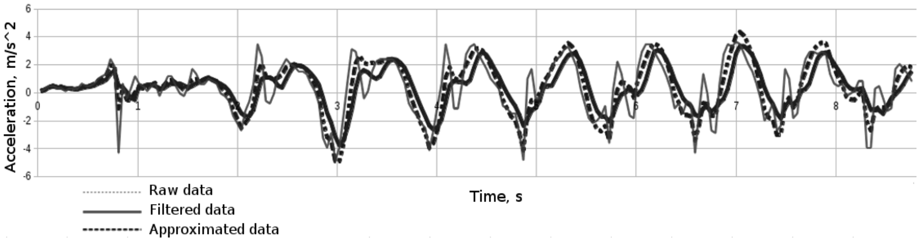


Fig. 3. The original time series, smoothed time series with the filter (n = 5), approximated time series by the least squares method (n = 7)

Results of smoothing and approximation are shown in Figure 3. It can be seen that on the sixth second, in the case of non-optimal choice of parameters of filtering, the wrong intersection of zero could be interpreted as additional step. In time points 0.8-1.0 (s) considerable oscillations were observed, which belong to process of fixing of the sensor on the body of the patient.

5 Minimizing the Objective Function

The mentioned formula for finding the MSE has been generalized for the cases of different values of parameters n (for smoothing) and m (for the approximation).

$$MSE = \frac{1}{N} \cdot \sum_{t=1}^{N} \left(\frac{1}{n} \cdot \sum_{i=0}^{n-1} x(t-i) - (a_t \cdot x(t) + b_t) \right)^2 \tag{4}$$

where N is the number of all the raw values, n is the number of previous raw values, a_t and b_t are the coefficients of linear approximation for m previously approximated values including the last raw value.

Since the objective function (MSE) is given in a discrete form, n and m take only integer values. Thus, in order to minimize such a function, discrete stochastic optimization techniques may be suitable, among which the following algorithms could be mentioned: simulated annealing, swarm part, and genetic selection.

Next, a comparison of these methods applied to the problem of minimizing MSE function of two arguments (n and m) is carried out.

6 Simulated Annealing Algorithm

The “simulated annealing” algorithm inherits the behavior of the atoms in the metal during heating and gradual cooling. The process of finding the optimal solution is controlled by the “metal temperature”. When the metal is heated up to the selected maximum temperature, the algorithm takes any possible solution as an optimal value.

The following relationship was used for calculating the probability of new solution acceptance [6] :

$$\Delta C = F(X_{current}) - F(X_{new}) \tag{5}$$

$$p = 1 \text{ for } \Delta C \le 0 \tag{6}$$

$$p = e^{\frac{-\Delta C}{T}} \text{ for } \Delta C > 0 \tag{7}$$

where F is the objective function (in our case, modified MSE), X are coordinates in the search space (n, m), p is the probability of adoption of a new decision, and T is the metal temperature.

During cooling, the confidence interval (expressed in terms of probability, as is given in (7)) is narrowing, making it difficult to find new solutions.

The current solution $X_{current}$ (a local minimum) is compared with the new value X_{new} (the vector is composed of random numbers). In the case of the positive difference (see (5)), X_{new} is considered as the current solution.

To reduce the possibility of stopping the optimization process at a local minimum, a comparison p with a random value p', which belongs to the [0, 1] interval, is conducted. In the case when the value of p' belongs to the [0; p] interval, X_{new} is taken as a current value.

Table 1. Parameters of the simulated annealing algorithm

The maximum number of iterations	100
The maximum temperature	10000.0
Temperature changing coefficient	0.99

The value of temperature coefficient was chosen to be equal to 0.99, for reduction of cooling speed, since temperature was estimated by the following relationship (see (7)): $T_{new} = T_{current} * 0.99$. In the case of lower coefficient values, during testing, quick temperature drop and stop on a local optimum were observed.

7 Particle Swarm Algorithm

The "particle swarm" algorithm inherits the social behavior of animals during the process of community activities: hunting, migration, and so on. Through the interaction, particles are able to find an optimal state in the swarm.

At the beginning, all the particles are distributed randomly in space, with different movement directions and speeds. At each iteration, a particle evaluates various position variants in space, which depend on the previous particle position, positions of the neighbors and the global position of swarm. The optimum speed and direction of motion are calculated for each particle at each iteration.

The equation for calculating the speed and direction of motion is as follows [7]:

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1(p_{ib} - x_i(t)) + c_2 \cdot r_2(p_{gb} - x_i(t)) \tag{8}$$

$$x_i(t+1) = x_i(t) + v_i(t) \tag{9}$$

where $v_i(t+1)$ is the new speed, w is the inertial weight, c_1 and c_2 are weights (constants) for the particle position and swarm position, r_1 and r_2 are random values [0; 1], p_{ib} and p_{gb} are values for the local optimal position of each particle and the position of a whole swarm, $x_i(t)$ is the position in space (at time t) for the certain particle.

Table 2. Parameters of the particle swarm algorithm

The maximum number of iterations	100
Number of particles	5
The maximum speed rate	3.0
Weight of particle position	1.0
Weight of general position	2.0

8 Genetic Selection Algorithm

The "genetic selection" algorithm inherits aspects of Darwin's theory of evolution by natural selection, simulating the process of gradual improvement of populations (candidate solutions) through the selection of the fittest individuals (solutions) which have, in our case, the less objective function value (the modified MSE)

In the evolution theory context, each individual (a candidate solution) is described by a group of genes (genome), and a set of individuals is called a population. The population size was defined by a constant P, and the genome of individuals consisted of two genes: m and n.

In order that the search would converge, the objective function value (the modified MSE) was evaluated for each individual, and individuals having the best values (their quantities were equal to constant E) were selected. The rest of the population (P - E) was obtained by mutation and crossover operators applied to the best individuals (E).

The mutation operator changes the value of a randomly chosen gene on the number in the selected limits. The crossover operator recombines genes of individuals [8]. In the current task, individual's genome consisted of two genes (n, m); in this case, the crossing was carried out by the cross replacement of single genes.

The choice between the mutation and crossover operators was implemented randomly; the crossover operator had the greater choice probability (0.7).

In order that the search would converge, the stopping criterion ($\varepsilon \leq MSE_{min}$) was introduced to the objective function, and the number of iterations without objective function (MSE) decreasing was limited. Search was repeated until the individual with the objective function value satisfying the stopping criterion was found or a number of iterations without changing exceeded the given value.

Table 3. Parameters of the genetic selection algorithm

The maximum number of iterations	51
The size of the genotype	2
Population size	15
Percentage of the best individuals keeping (elitism rate)	0.3
The probability of crossing / mutation	0.7/0.3

9 Results of Testing

As a result of minimization of the objective function (MSE), the best parameter values for the moving average filter were found.

Table 4. Optimal values of n for speeds

Speed, m/s	0.926	1.032	1.316
n (for smoothing) / m (for approximation)	10/10	9/9	12/11

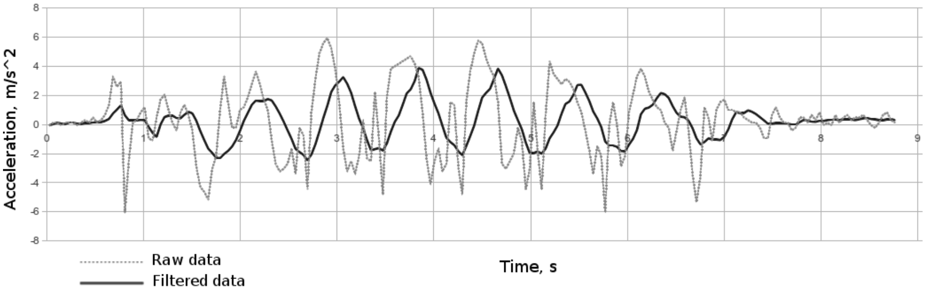


Fig. 4. Original time series for the speed of 1.032 (m/s), (6 sample steps), smoothed values

The results of the optimal data filtration are shown in Figure 4.

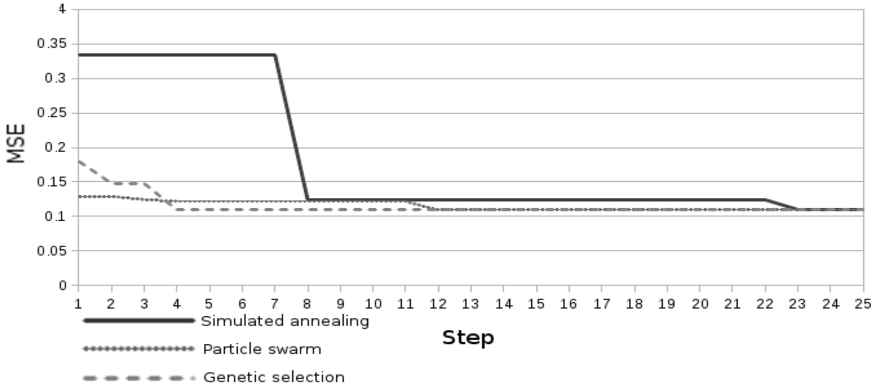


Fig. 5. Comparison of convergence of algorithms, for speed of 0.926 (m/s)

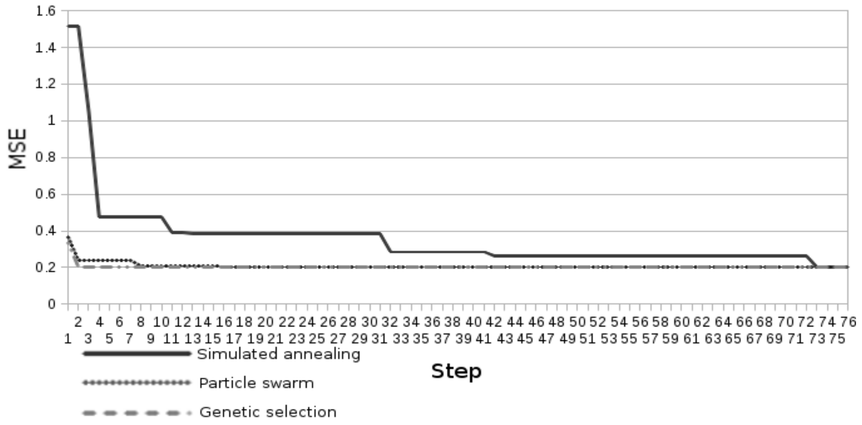


Fig. 6. Comparison of convergence of algorithms, for speed of 1.032 (m/s)

Table 5. Convergence of algorithms

Speed, m/s	0.926	1.032	1.316
	Number of iterations		
Simulated annealing	23	72	48
Particle swarm	12	15	18
Genetic selection	4	2	7

The results of the convergence of the algorithms (simulated annealing, particle swarm and genetic selection algorithms) are given in Table 5 and Figures 5, 6.

10 Conclusions

The reason for writing this article was the study of the realization of the adaptive filters for human movement registration systems (walking was considered). When walking with different speeds, the process behavior characteristics registered by accelerometer varied widely.

Because setting the optimal filter parameters for each case is not possible, it is necessary to have a system capable for learning during the process of walking. Training includes the search for optimal filtering options for different types of movement, clustering movements and memorizing their parameters.

For comparison of stochastic optimization methods, the math library has been created, which included the following methods: linear approximation, calculation of the mean square error, standard moving average filtration, the simulated annealing algorithm, the particle swarm algorithm, and the genetic selection algorithm.

During selection of the optimization algorithm, among the main features, processing speed and minimal consumption of resources should be considered. The appropriate choice of parameters for the particle swarm algorithm, such as the number of particles and the weights of the particle and swarm positions, allows us to achieve the same optimization quality as in the genetic selection algorithm. Among the stochastic optimization methods, the simulated annealing algorithm exhibited the most time to find the optimum because of slow convergence to the global optimum.

As our experiments showed, the algorithms mentioned above, as well as the library created can be used for the effective adaptive calculation of the low-pass filter parameters.

References

1. Yang, G.Z.: Body Sensor Networks, pp. 1–10. Springer, London (2006)
2. Bourke, A., O'Brien, J., Lyons, G.: Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. Gait and Posture 26, 194–199 (2007)
3. Matsumura, T.: Device for Measuring Real-time Energy Expenditure by Heart Rate and Acceleration for Diabetic. In: Patients, T., Matsumura, V.T., Chemmalil, M.L., Gray, J.E., Keating, R.L. (eds.) 35th Annual Northeast Bioengineering Conference, Boston, pp. 1–2 (2009)
4. Wang, D.: Compared performances of morphological, median type and running mean filters. In: Wang, D., Ronsin, J., Haese-Coat, V. (eds.) Visual Communications and Image Processing. SPIE, vol. 1818, pp. 384–391 (1992)
5. Ng, L.: Fast moving average recursive least mean square fit. In: Ng, L., LaTourette, R. (eds.) 24th Conference on Decision and Control, pp. 1635–1636 (1985)
6. Vicentea, J., Lancharesb, J., Hermida, R.: Placement by thermodynamic simulated annealing. Physics Letters A 317(5-6), 415–423 (2003)
7. Parsopoulos, E.: Particle Swarm Optimization Method in Multiobjective Problems. In: Parsopoules, E., Vrahatis, N. (eds.) Symposium on Applied Computing, pp. 603–607 (2002)
8. Bessaou, M., Siarry, P.: A genetic algorithm with real-value coding to optimize multimodal continuous functions. Structural and Multidisciplinary Optimization 23, 63–74 (2001)

The Performance of Concatenated Schemes Based on Non-binary Multithreshold Decoders

Valery Zolotarev[1], Gennady Ovechkin[2], Pavel Ovechkin[2], Dina Satibaldina[3], and Nurlan Tashatov[3]

[1] Space Research Institute of Russian Science Academy, Moscow, Russian Federation
zolotasd@yandex.ru
[2] Ryazan State Radioengineering University, Ryazan, Russian Federation
{g_ovechkin,pavel_ov}@mail.ru
[3] L.N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan
satybaldina_dzh@enu.kz, tash.nur@mail.ru

Abstract. Symbolic *q*-ary multithreshold decoding (*q*MTD) for *q*-ary self-orthogonal codes (*q*SOC) is analyzed. The SER performance of *q*MTD is shown to be close to the results provided by optimum total search methods, which are not realizable for non-binary codes in general. *q*MTD decoders are compared with different decoders for Reed-Solomon and LDPC codes. The results of concatenation of *q*SOC with simple to decode outer codes are described. The complexity of *q*MTD is also discussed.

Keywords: iterative decoding, non-binary (symbolic) multithreshold decoding, *q*-ary self-orthogonal codes, concatenated codes, symbolic codes.

1 Introduction

Error correcting coding is used to correct errors appearing during data transmission via channels with noises. Main attention in literature is given to binary error-correcting codes working with data on the level of separate bits. But in many digital systems it's often more convenient to work with byte structure data. As an example, it's more convenient to work with bytes in systems which store big volumes of data (optic discs and other devices). In such systems to protect data from errors it is recommended to use non-binary error-correcting codes. At present preference among non-binary codes is given to Reed-Solomod codes (RS), which have algebraic decoding algorithms [1], allowing to correct up to half-distance errors as well as more complex algorithms [2] providing correction of higher error number. At the same time due to their implementation complexity such methods allow to decode only short and thus low-effective RS codes. Lately many specialists have been developing decoders of non-binary low-density parity-check (*q*LDPC) codes which are able to provide extremely high efficiency [3, 4]. But the complexity of such decoders especially with big alphabet size still remains too high to be used in practice.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_24, © Springer International Publishing Switzerland 2014

Special attention among non-binary algorithms to correct errors should be given to non-binary (q-ary) self-orthogonal codes and special high-speed alphabet multithreshold decoders (qMTD) corresponding to them [5…8], being the development of binary multithreshold decoders (MTD) [5, 6, 9…10]. Great interest to MTD is shown not only in Russia [11, 12]. Research results given in [5…8] show that qMTD greatly exceed in their efficiency RS codes and qLDPC codes being used in practice remaining as simple to be implemented as their prototypes – binary MTD**.** It is also very important not to use multiplication in non-binary fields during encoding and decoding as well as total independence of alphabet codes lengths from the size of applied symbols. That's the reason why such codes will find broad application in the sphere of processing, storage and transmission of large volumes of audio, video and other types of data.

The rest part of the article is organized in the following way. Section 2 contains basic information about qMTD. Section 3 shows the results of qMTD efficiency comparison with efficiency of decoders for RS and qLDPC codes. Section 4 is dedicated to the development of new concatenated schemes to correct errors based on qMTD and their efficiency analysis. Section 5 demonstrates the main conclusions.

2 Non-binary Multithreshold Decoding

Let's describe operating principles of qMTD during non-binary self-orthogonal codes (qSOC) decoding. The description is given for q-ary symmetric channel (qSC) having alphabet size q, $q > 2$, and symbol error probability p_0.

Let's assume linear non-binary systematic convolutional or block self-orthogonal code with parity-check matrix **H** to be equal to binary case [6, 13], i.e. it has only zeros and ones excluding the fact that instead of 1 there will be –1 in identity submatrix, i.e. $\mathbf{H} = [\mathbf{P} : -\mathbf{I}]$. Here **P –** submatrix defined by generator polynomial for binary SOC; **I** – identity submatrix. Generator matrix of such code will be of $\mathbf{G} = [\mathbf{I} : \mathbf{P}^{\mathrm{T}}]$ type. This code can be used with any size q of alphabet.

Note that for this qSOC during encoding and decoding operations only addition and subtraction on q module are necessary to be made. Calculations in non-binary fields are not applied in this case.

The example of a scheme realizing the operation of encoding by block qSOC, given by generator polynomial $g(x) = 1+x+x^4+x^6$, is given on Fig. 1. Such code is characterised by the parameters: code length n=26 symbols, data part length k=13 symbols, code rate R=1/2, code distance d=5.

Let's assume that encoder has performed encoding of data vector $\boldsymbol{U}$ and received code vector $\boldsymbol{A} = [\boldsymbol{U}, \boldsymbol{V}]$, where $\boldsymbol{V} = \boldsymbol{U} \cdot \mathbf{G}$. Note that in this example and below when multiplication, addition, subtraction of vectors and matrices are made, module arithmetics is applied. When code vector $\boldsymbol{A}$ having the length n with k data symbols on qSC is transmitted decoder is entered with vector $\boldsymbol{Q}$, generally speaking, having differences from original code vector due to errors in the channel: $\boldsymbol{Q} = \boldsymbol{A} + \boldsymbol{E}$, where $\boldsymbol{E}$ – channel error vector of qSC type.

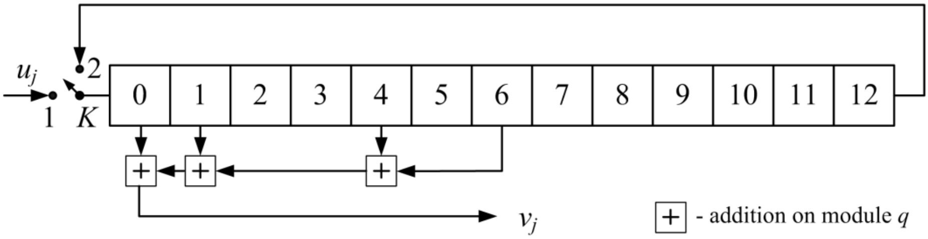


Fig. 1. Encoder for block qSOC, given by polynomial $g(x) = 1+x+x^4+x^6$

Operating algorithm of qMTD during vector $\boldsymbol{Q}$ decoding is the following [5…8].

1. Syndrome vector is calculated $\boldsymbol{S} = \mathbf{H} \cdot \boldsymbol{Q}^{\mathrm{T}}$. Difference register $\boldsymbol{D}$ is reset. This register will contain data symbols changed by decoder. Note that the number of non-zero elements of $\boldsymbol{D}$ and $\boldsymbol{S}$ vectors will always determine the distance between message $\boldsymbol{Q}$ received from the channel and code word being the current solution of qMTD. The task of decoder is to find such code word which demands minimal number of non-zero elements of $\boldsymbol{D}$ and $\boldsymbol{S}$ vectors. This step totally corresponds to binary case.

2. For arbitrarily chosen decoded q-ary data symbol i_j of the received message let's count the number of two most frequent values of checks s_j of syndrome vector $\boldsymbol{S}$ from total number of all checks relating to symbol i_j, and symbol d_j of $\boldsymbol{D}$ vector, corresponding to i_j symbol. Let the values of these two checks be equal to h_0 and h_1, and their number be equal to m_0 and m_1 correspondingly when $m_0 \geq m_1$. This step is an analogue of sum reception procedure on a threshold element in binary MTD.

3. If $m_0 – m_1 > T$, where T – a value of a threshold (some integer number), then from i_j, d_j and all checks regarding i_j error estimation equal to h_0 is subtracted. This step is analogous to comparison of a sum with a threshold in binary decoder and change of decoded symbol and correction via feedback of all syndrome symbols being the checks for decoded symbol.

4. The choice of new i_m, $m \neq j$ is made, next step is clause 2.

Such attempts of decoding according to cl. 2…4 can be repeated for each symbol of received message several times [5, 6]. Note that when implementing qMTD algorithm the same as in binary case it is convenient to change all data symbols consequently and to stop decoding procedure after fixed number of error correction attempts (iterations) or if during such iteration no symbol changed its value. The example of qMTD implementation for encoder from Fig. 1 is given on Fig. 2.

For qMTD algorithm described the following theorem is valid,

Theorem. Let decoder realize qMTD algorithm for the code described above. Then during each change of decoded symbols a transition to a more plausible solution in comparison with previous decoder solutions takes place.

Proof of a theorem is given in [5…8]. In the course of proof it is shown that total Hamming weight of syndrome and difference registers during each change of decoded symbols in accordance with qMTD algorithm described above strictly decreases.

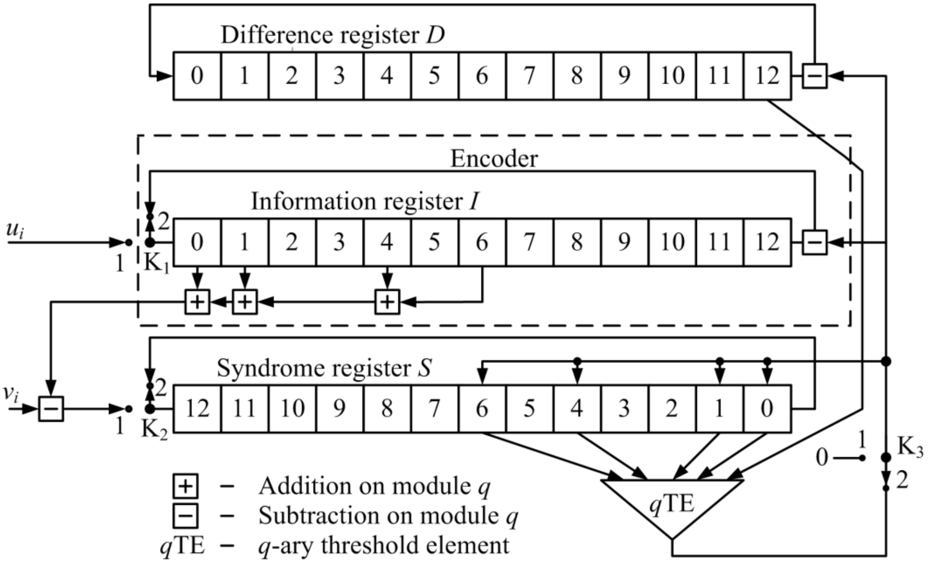


Fig. 2. MTD for block *q*SOC

Let's note two most important features characterizing offered algorithm. First, as in case of binary codes we can't claim that *q*MTD solution improvement during multiple decoding attempts will take place till optimal decoder solution is achieved. In fact both in block and in convolutional codes it's possible to meet such error configurations which cannot be corrected in *q*MTD, but some of them can be corrected in optimal decoder. That's why the main way to increase *q*MTD efficiency is to search codes where these noncorrected error configurations are quite rare even in high level of noise. The questions to choose such codes are considered in detail in [6].

Another important moment is the fact that in comparison with traditional approach to major systems to change decoded symbol *q*MTD needs not absolute but relatively strict majority of checks as it follows from m_0–$m_1 > T$ condition. E.g., in *q*SOC with $d = 9$ an error in decoded symbol will be corrected even when only 2 checks will be correct from 9 his checks (including symbol d_j of difference register) and the other 7 - erroneous! This situation cannot be imagined for binary codes but for *q*MTD this is typical.

These features essentially expand the possibilities of non-binary multithreshold algorithm during operation in high noises retaining as it follows from given description only linear dependence of implementation complexity from code length.

3 Simulation Results

Let's compare characteristics of *q*MTD and other non-binary error correction methods in *q*SC. Dependencies of symbol error rate P_s after decoding from symbol error P_0 probability in *q*SC for codes with code rate R=1/2 are given in Fig. 3. Here curves *5* and *6* show characteristics of *q*MTD for *q*SOC with block length n=4000 and 32000 symbols when using 8-bit symbols (alphabet size q=256). The volume of simulation in lower points of these graphs contained from $5\cdot10^{10}$ to $2\cdot10^{12}$ symbols which shows extreme method simplicity. As a comparison in this Figure curve *1* shows characteristics of algebraic decoder for (255, 128) RS code for q=256. As it follows from the Fig. 3, efficiency of qMTD for qSOC turns out to be far better than efficiency of RS code decoders using the symbols of similar size. When code length in qMTD increases the difference in efficiency turns out to be even higher. Note that even when using concatenated schemes of error correction based on RS codes it's not possible to increase decoding efficiency considerably. E.g., with the help of product-code having code rate 1/2, consisting of two RS codes with q=256 and several dozens of decoding iterations error rate less than 10^{-5} can be provided with error probability in the channel only equal to 0,18 (curve *4* in Fig. 3), which is considerably worse than when using qMTD. Besides different methods to increase correcting capability of RS codes including all variations of Sudan algorithm ideally have the complexity of n^2 order. For the codes having the length of 32000 symbols this leads to the difference in complexity equal to 32000 times having at the same time little increase of error-correctness. This is shown in Fig. 3 with curve *3*, which gives the estimations of Wu [2] algorithm possibilities for (255, 128) of RS code.

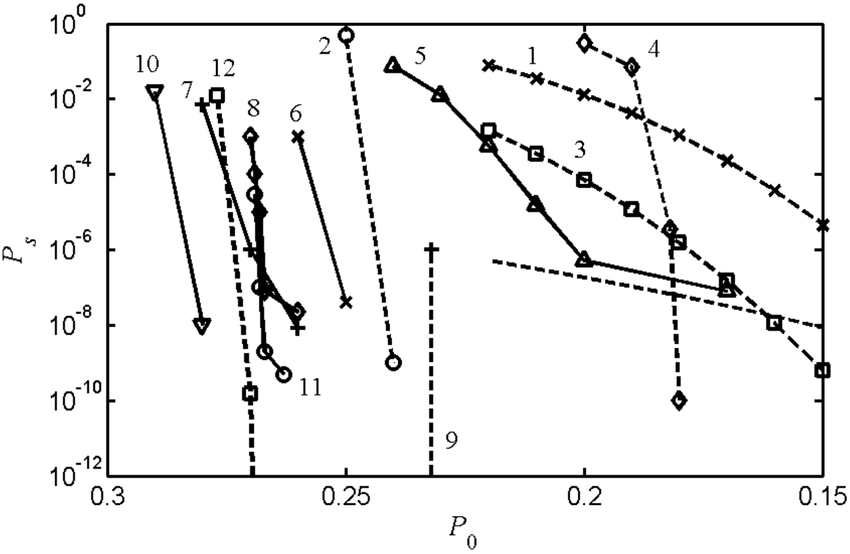


Fig. 3. Characteristics of non-binary codes with code rate R=1/2 in qSC

Additional advantage of qMTD over other error correcting methods is the fact that it allows to work easily with symbols of any size providing high correcting capability. This is confirmed by curves which show characteristics of qMTD for code having the length equal to 32000 two-byte symbols (curve *7*) and to 100000 four-byte symbols (curve *10*). We should note that very simple to be implemented qMTD decoder for two-byte code with the length 32000 is capable to provide error-correctness not accessible even by RS code with the length of 65535 two-byte symbols (curve *2* in Fig. 3), the decoder for which is not to be implemented in close future. Besides, *q*MTD for four-byte symbols even surpasses in efficiency more complicated decoder of *q*LDPC codes with the length of 100000 four-byte symbols which has the example of characteristics presented in Fig. 3 by curve *9* [4].

It should be noted once more that to achieve these results with the help of *q*MTD used codes should be chosen very thoroughly and the main criterion while choosing should be the degree of resistance to the effect of error propagation. At the same time the most effective are the codes where several data and several check branches are used [6, 14]. In [15] it is shown that in the process of such codes optimizing it is possible to improve *q*MTD operating efficiency. Particularly, characteristics of the code with q=256 and code rate 1/2 found in [15] are given in Fig. 3 by curve *8*. It is clear that this code provides effective work in conditions of bigger error probabilities in *q*CK, than the codes known before (curve *6*), having the same complexity of their decoding.

4 Concatenated Schemes of Error Correction Based on *q*MTD

One of the ways to improve *q*MTD characteristics is to use it in concatenated encoding schemes. The simplest and most effective concatenated encoding scheme is concatenated scheme on the basis of *q*SOC and control code on module q [6, 8, 16]. In the field of its effective work *q*MTD is known to leave only rare single errors. The task to correct such single errors is easily solved with the help of control codes on module q.

The process of encoding by concatenated code encoder on the basis of *q*SOC and control code on module q is the following. First each sequence consisting of n–1 symbols is complemented by such n-th symbol that the sum of symbols value having the sequence of n elements on q module becomes equal to 0. After that this new sequence of n elements is encoded by *q*SOC encoder. Decoding process of the message received from the channel is made in reverse order, i.e. non-binary multithreshold decoding is made first after which in the conditions of lower noise level channel contains basically single errors which are corrected by decoder for control code on module q.

Operation of decoder for control code on module q is the following. First e sum on module q values for block consisting of n elements is calculated. If this sum is not equal to 0, then among the first n–1 elements in the block the one with less reliability should be found, the reliability of which is less than reliability of n-th symbol in block. If such symbol exists then it is changed on e value. The reliability here is understood as m_0–m_e difference, where m_0 – number of zero symbols of syndrome and difference register of *q*MTD connected with given data symbol; m_e – number of symbols of syndrome and difference register of *q*MTD with the value e and connected with given symbol.

In Fig. 3 curve *11* shows characteristics of concatenated encoding scheme consisting of *q*SOC and control code on module q in *q*SC. Inner code was *q*SOC with minimum code distance d=17 and code rate R=8/16 the characteristics of which are represented by curve *8*. Outer code was control code on module q with the length L=50. During qSOC decoding qMTD with 30 iterations was used. The Figure shows that usage of decoder for control code on module q with block length L=50 after qMTD allows to reduce decoding error rate on more than two orders. The increase of calculation volume in concatenated code is less than 20% in comparison with original qMTD algorithm.

Essential drawback of the concatenated scheme described above is the fact that decoder of outer control code on module q sometimes does not correct even the only error in the block. To eliminate this drawback it is recommended to use together with qSOC more effective and simple to be implemented non-binary code the decoder of which will always correct the only symbol error in the block. This will allow to reduce error rate in the field of effective qMTD operation even more in comparison with concatenated scheme presented above. As an example of such code non-binary Hamming codes [17] can be used. At the same time known non-binary Hamming codes have such features as the necessity to use extended Galois fields in the process of decoding as well as dependence of code length from alphabet size. As a result the application of such codes in offered concatenated scheme especially when the alphabet is big becomes too complicated. That's why it could be offered to build non-binary Hamming codes [16] on the basis of known binary Hamming codes. Let's describe them in detail.

Parity-check matrix of these codes coincides with parity-check matrix of binary Hamming codes $\mathbf{H}_h$=[$\mathbf{C}_h$: $\mathbf{I}$]. Generator matrix will be as follows $\mathbf{G}_h$=[$\mathbf{I}$: -$\mathbf{C}_h^T$]. Let us formulate the principles to decode this code.

Let's assume that after qMTD vector $\boldsymbol{Y}$ entered input of non-binary Hamming code decoder. In the process of decoding a syndrome of received message is calculated first:

$$\boldsymbol{S}_h = \boldsymbol{Y} \cdot \mathbf{H}_h^T.$$

If the received message contains only one error with value e_j on j position then generated syndrome can be written down as

$$\boldsymbol{S}_h = \boldsymbol{S}_2^j e_j,$$

where $\boldsymbol{S}_2^j$ – syndrome of binary Hamming code with single error on j position. Consequently, such symbol of received message need to be corrected on value e_j for which a column of parity-check matrix $\mathbf{H}_h$ coincides with syndrom $\boldsymbol{S}_2^j$.

If received message contains two errors e_i and e_j on i and j positions then the syndrom can be written down as follows

$$\boldsymbol{S}_h = \boldsymbol{S}_2^i e_i + \boldsymbol{S}_2^j e_j.$$

Such syndrome contains only values 0, e_i, e_j and e_i+e_j. Consequently, such symbol of received message need to be corrected on value e_i for which matrix column $\mathbf{H}_h$ coincides with vector $\boldsymbol{S}_2^i$, and such symbol of received message need to be corrected on value e_j for which matrix column $\mathbf{H}_h$ coincides with vector $\boldsymbol{S}_2^j$. Thus, offered algorithm of offered non-binary Hamming codes decoding in majority of cases (approximately in 71% cases for q=256 [16]) is able to correct even two errors. And if it is use offered extended non-binary Hamming codes having in addition one general check on module q then two errors are practically corrected in all cases (in 99% of cases for q=256 [16]).

The example of performance of offered concatenated scheme containing qSOC with R=8/16, q=256, d=17 and given extended non-binary Hamming code with the length N_2=128 is shown in Fig. 3 by curve *12*. At the same time total decoding complexity due to addition of extended non-binary Hamming code increases not more than 35 % [16]).

5 Conclusion

Given results allow to conclude that qMTD methods can really be regarded as unique algorithms capable to provide effective decoding in the conditions of high noise level requaring quite small number of operations and achieving highest levels of reliability in the process of digital information transmission and storage as well as its processing rate in high-speed communication channels and in the devices to store large volume of data.

The work was supported by Russian foundation for basic research, Space research institute, Ryazan state radioengineering university and Science comitte of Republic of Kazakhstan. Great deal of additional information on multithreshold decoders can be found on websites [18].

References

1. Berlekamp, E.R.: Algebraic Coding Theory. McGraw-Hill, New York (1968)
2. Wu, C.: New list decoding algorithms for Reed-Solomon and BCH codes. IEEE Transactions on Information Theory 54, 3611–3630 (2008)
3. Declercq, D., Fossorier, M.: Extended minsum algorithm for decoding LDPC codes over GF(q). In: IEEE International Symp. on Inf. Theory, pp. 464–468 (2005)
4. Zhang, F., Pfister, H.: List-Message Passing Achieves Capacity on the q-ary Symmetric Channel for Large q. In: Proc. IEEE Global Telecom. Conf., Washington, DC, pp. 283–287 (November 2007)
5. Zubarev, U.B., Zolotarev, V.V., Ovechkin, G.V.: Review of error-correcting coding methods with use of multithreshold decoders. Digital Signal Processing (1), 2–11 (2008)
6. Zolotarev, V.V., Zubarev, Y.B., Ovechkin G.V.: Multithreshold decoders and optimization coding theory. In: Hot line – Telecom, 239 p. (2012)
7. Zolotarev, V.V., Averin, S.V.: Non-Binary Multithreshold Decoders with Almost Optimal Performance. In: 9th ISCTA 2007, UK, Ambleside (July 2007)
8. Ovechkin, G.V., Zolotarev, V.V.: Non-binary multithreshold decoders of symbolic self-orthogonal codes for q-ary symmetric channels. In: 11th ISCTA 2009, UK, Ambleside (July 2009)
9. Zolotarev, V.V., Ovechkin, G.V.: The algorithm of multithreshold decoding for Gaussian channels. Information Processes 8(1), 68–93 (2008)

10. Ovechkin, G.V., Zolotarev, V.V., Averin, S.V.: Algorithm of multithreshold decoding for self-orthogonal codes over Gaussian channels. In: 11th ISCTA 2009, UK, Ambleside (July 2009)
11. Ullah, M.A., Okada, K., Ogivara, H.: Multi-Stage Threshold Decoding for Self-Orthogonal Convolutional Codes. IEICE Trans. Fundamentals E93-A(11), 1932–1941 (2010)
12. Ullah, M.A., Omura, R., Sato, T., Ogivara, H.: Multi-Stage Threshold Decoding for High Rate Convolutional Codes for Optical Communications. In: AICT 2011: The Seventh Advanced international Conference on Telecommunications, pp. 87–93 (2011)
13. Massey, J.: Threshold decoding. M.I.T. Press, Cambridge (1963)
14. Davydov, A.A., Zolotarev, V.V., Samoilenko, S.I., Tretiakova, Y.I.: Computer networks. – M.: Science (1981)
15. Ovechkin, G.V., Ovechkin, P.V.: Optimisation of non-binary self-orthogonal codes structure for parallel coding schemes. In: NIIR FSUE, vol. (2), pp. 34–38 (2009)
16. Ovechkin, G.V., Ovechkin, P.V.: Usage of non-binary multithreshold decoder in concatenated shemes of errors correction. RSREU Journal, №4 (issue 30) (2009)
17. Ling, S., Xing, C.: Coding theory. A first course, Cambridge (2004)
18. Web sites of IKI `http://www.mtdbest.iki.rssi.ru` and RSREU `http://www.mtdbest.ru`

Novel Precoded OFDM for Cognitive Radio in Noisy and Multipath Channels

Sundaram K. Padmanabhan[1] and T. Jayachandra Prasad[2]

[1] Sathyabama University, Chennai
padusk1@gmail.com
[2] Rajeev Gandhi Memorial College of Engineering and Technology, Andra Pradesh
jp.talari@gmail.com

Abstract. The wireless spectrum being a scanty resource needs to cater up for stipulated number of devices. But with wireless devices becoming pervasive, the provision of spectrum for such a substantial number seems to be a daunting task. Cognitive radio is a technology which uses the spectrum smartly either by sensing the spectrum hole in order to transmit the data or by sharing the available spectrum among two or more users. Hence cognitive radio is indispensible for the spectrum prerequisites. But the main problem is the phenomenon of cognitive interference. Precoded OFDM (Orthogonal Frequency Division Multiplexing) is found to eliminate the interference by such a way that it gets cancelled as it is propagated. The precoded OFDM is similar to that of the VFDM system except for the fact that the secondary receiver does not requires to know the channel state information but only in which mode the transmitter is being operated. This work comprises of the analysis of BER (Bit Error Rate) performance of the Precoded system in AWGN (Additive White Gaussian Noise), Rayleigh and Rician channels using MATLAB.

Keywords: OFDM, cognitive radio, dynamic spectrum usage, overlay networks, precoding.

1 Introduction

Cognitive radio is a technology which is current burgeon area of research proves to solve the glitch of unavailability of spectrum. The cognitive radio is based on dynamically allocating the spectrum by recognizing the unused spectrum. Even though the spectrum can be easily perceived through the usage of blind spectrum estimation techniques, the transmission without causing interference to other counter parts seems to be strenuous. Hence the system has to transmit messages without causing interference considering that the other system is entirely using the same spectrum. Consider a scenario where a user is currently occupying and transmitting messages using a spectrum K. This system is known as primary system while the transmitter and receiver are known as the primary transmitter and receiver respectively. Consider a case that a new system enters the cell and finds itself in a position of ideal state due to unavailability of spectrum. In such a situation instead of waiting for the spectrum to be

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_25, © Springer International Publishing Switzerland 2014

freed, provision can be made in order to transmit the message over the spectrum of the primary. This system which shares the primary's spectrum is known as the secondary system. Its transmitter is called as the secondary transmitter while the receiver is known as the secondary receiver. For such interference free transmission the secondary transmitter has to know the primary system message [2]. This can become difficult as the primary could not even have idea about the presence of such secondary system [3]. Also in non-contiguous OFDM transmission of messages the primary and the secondary system has to perfectly synchronized in order to eliminate interference so that either of the system knows which are subcarriers to be used [4].

The Vondermonde Frequency Division Multiplexing (VFDM) scheme proposed in [5-7] is a propitious system which efficiently eliminates the interference at the primary receiver by precoding the message with a Vondermonde matrix. The major disadvantage is that the system requires the secondary to know the channel matrix [6].

The main objective of this paper is to propose a method by which the secondary system's message can be transmitted using the primary's spectrum without causing any sort of interference to the primary. Also the proposed system is evaluated for BER (Bit Error Rate) in varying channels.

2 OFDM

This section briefs about the OFDM system. OFDM is a multicarrier modulation technique which uses N sub carriers to transmit messages. In its initial architecture it consisted of N local oscillator which made it complex to implement. Hence a digital implementation of the subcarriers is carried out by using an IFFT (Inverse Fast Fourier Transform) structure.

IFFT is given by the following equation

$$X(k) = \sum_{x=0}^{n-1} x(n).e^{-\frac{j2\pi nk}{N}} \tag{1}$$

Where k=0, …, N-1, where N represents the number of subcarriers. The OFDM transmitter is shown in figure 1.

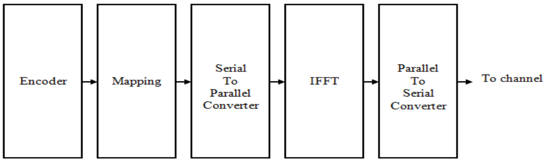


Fig. 1. OFDM Transmitter

The data is first encoded and then mapped to symbols. Once the mapping is carried out the serial data is converted to a parallel data which is then inverse Fourier transformed. Generally the Fourier transform is employed as butterfly structure, which enhances the speed of operation. Hence they are implemented as Inverse Fast Fourier Transform.

After the IFFT the data are then converted back as serial data. The serial data is then transmitted through the channel [3].

The receiver performs the inverse operation of transmitter. Once the data has been reached at the receiver the serial data is converted to parallel data. The Fast Fourier transform is carried out which is the inverse of the IFFT operation. After FFT the parallel data is converted back to serial data. The message is demapped in order to convert symbol back to bits. The bits are then decoded in order to recover back the message bits. The following figure depicts the OFDM receiver operation.

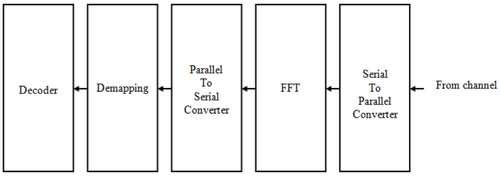


Fig. 2. OFDM Receiver

Even though the conventional OFDM system cannot be used for serving cognitive radio prerequisites, they serve as basis for system involving cognitive radio

3 VFDM System

The Vondermonde Frequency Division Multiplexing eliminates the interference at the primary caused by the secondary. Consider a cognitive channel with a primary and a receiver shown below. The primary transmits its message through its direct channel h_{11} and secondary through h_{22} . They cause interference to each other through their indirect channel h_{12} and h_{21} .

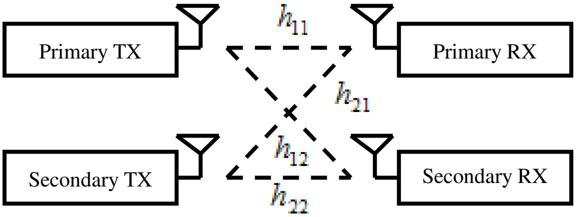


Fig. 3. Cognitive Interference Channel

Hence the message signals which are received at the primary receivers can be represented mathematically as

$$\Re_1 = H(h_{11})x_1 + H(h_{21})x_2 + n_1 \tag{2}$$

And at the secondary receiver it is given as,

$$\Re_2 = H(h_{22})x_2 + H(h_{12})x_1 + n_2 \tag{3}$$

Where the terms n_1 and n_2 represents the noise from other sources in the channel. In the equation (2) the notation $H(h_{ij})$ represents the channel between the i_{th} transmitter and the j_{th} receiver which modeled from the channel filter tap weights [8] as a Toeplitz matrix of order $(L+L_p)\times(L+L_p+L_c-1)$

$$H(h_{ij}) = \begin{bmatrix} h_1^{ij} & \dots & h_{L_c}^{ij} & 0 & \dots & \dots & \dots & 0 \\ 0 & h_1^{ij} & \dots & h_{L_c}^{ij} & 0 & \dots & \dots & 0 \\ \vdots & & \ddots & & \ddots & & & \vdots \\ \vdots & \dots & 0 & h_1^{ij} & \dots & h_{L_c}^{ij} & 0 & 0 \\ \vdots & & & \ddots & \ddots & & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & h_1^{ij} & \dots & h_{L_c}^{ij} \end{bmatrix} \tag{4}$$

In the above relation h_m^{ij} represents the channel filter tap for $i = 1, \dots, L_c$, L_c being the number of multipath components for the channel h_{ij}. The channels are considered to i.i.d (i.e. no two channels will have same value of $H(h_{ij})$).

The main problem at the primary receiver is the interference caused by the secondary channel through its indirect channel h_{21} . Hence for zeroing this effect, the message has to be transmitted in such a way that it satisfies the following equation.

$$H(h_{21})x_2 = 0 \tag{5}$$

So that the message received at the primary receiver becomes

$$\Re_1 = H(h_{11})x_1 + n_1 \tag{6}$$

Hence in VFDM system, the process carried out to eliminate the secondary interference is to multiply a Vondermonde matrix V so that the following Equation is satisfied.

$$H(h_{22})Vx_2 = 0 \tag{7}$$

There value V is obtained as a linear precoder value which is given by

$$V = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_L \\ a_1^2 & \cdots & a_L^2 \\ \vdots & \ddots & \vdots \\ a_1^{N+L-1} & \cdots & a_L^{N+L-1} \end{bmatrix} \tag{8}$$

Where, the terms $a_1, \ldots, a_L$ are the roots of the polynomial $S(z) = \sum_{i=0}^{L} h_i^{(21)} z^{L-i}$.

As already discussed even though the system eliminates the interference it is mandatory that the secondary receiver must have knowledge about the channel for which the transmitter is precoding.

4 Precoded OFDM System

The precoded OFDM proposed in this work , is similar to that of the VFDM system except for the fact that the secondary receiver does not requires to know the channel state information but only in which mode the transmitter is being operated. The convention used in the work is that if the system works as a primary then the operating mode is 1. In case if the system is operated as secondary then the system is said to have an operating mode of 2. In order to make the primary interference free, the values have to be precoded in such a way that they satisfy equation (5). But doing so would cause the message recovery at the secondary impossible with the values which has been precoded. Hence a property of IFFT is used which is nothing but, if the second half of input of IFFT is given as repeat of first half then the output of IFFT will be zero at every even places. Mathematically it can be given as

$$F^{-1}(x(n))\,|_{x(n)=x\left(n+\frac{N}{2}\right)} => X(2k) = 0 \tag{9}$$

Thus at zeros places the values can inserted so that it satisfies (5). Such values are called as root values. At receiver if the system is found to be operating as secondary then it removes the values at even position and carries out the conventional OFDM process.

The transmitter block is shown in above figure. The operating mode is responsible for intimating the other blocks whether the current operation is carried out for transmitting messages in primary or in secondary mode. The operating mode is 1 when the system has dedicated spectrum i.e. when the system is operating as primary. When the system is operating as secondary then the mode is said to be 2. The next block to the encoder is the constellation mapping which is used to map bits to symbols. The serial data is converted to parallel data if the operating mode is 1. The block next to serial to parallel conversion is the buffer which in operating mode 1 performs no operation while being secondary it stores certain values without which when propagated requires a 2N point IFFT as there is a additional bit insertion in the form of roots. Hence the buffer ensures that only N symbols are propagated to the IFFT block which is same as that of the operating mode 1.

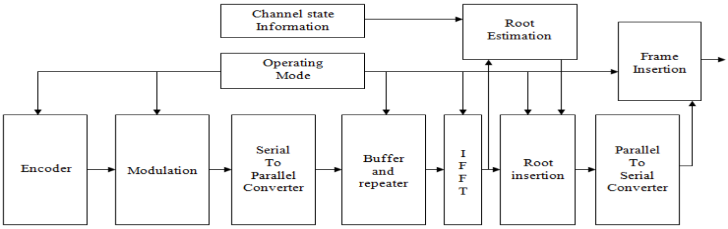


Fig. 4. Precoded OFDM transmitter

In case the operating mode is 2 then the parallel data is repeated to satisfy the property of IFFT which is mentioned in (9). In an OFDM system the mapped symbols are inverse Fourier transformed which is nothing but a digital implementation of subcarrier multiplication [3]. The Fourier transformed outputs are then sent for root estimation, where from the channel state information available, the roots are estimated so the propagation of messages along with the indirect secondary channel are eliminated. Once the roots are estimated then the values are inserted at even position where the IFFT outputs are zero. The frame bits are added in order to intimate the secondary about the operating mode of transmitter to the receiver. The frame consists of the data followed by mode field which specifies the current operating mode of the transmitter. The EXT field consist of additional information which is necessary for transfer control.

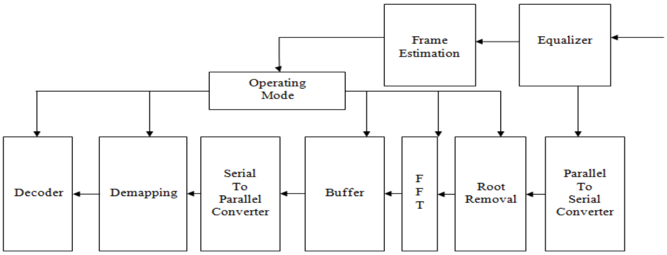


Fig. 5. Precoded OFDM Receiver

At the receiver, the frame is removed and is processed by frame estimator which then sends the control signal to the buffer and the adaptive demodulation block. When the frame estimator finds that the mode is operated under cognitive channel condition it then intimate the buffer about the mode. The FFT process the data as normal after the root removal, which removes the values present at even samples. The buffer then stores the data and forwards upon receiving the next N/2samples. The data is then subjected to demapping which then provides the information as output.

5 Root Estimation

This section describes the process of estimating the roots which is necessary to eliminate the interference. The assumption is made that the transmitter knows the channel state information perfectly. In a general case the output of the IFFT can be given as follows,

$$F^{-1}(x(n)) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_{N-1} \\ x_N \end{bmatrix} \tag{10}$$

Where, $F^{-1}(x(n))$ represents the inverse Fast Fourier transform on input x (n). As specified in the equation no (8) in the previous section the values are repeated so that the even position of the output of IFFT becomes zero which is shown in the following equation.

$$F^{-1}(x(n))\,|_{x(n)=x\left(n+\frac{N}{2}\right)} = \begin{bmatrix} x_1 \\ 0 \\ x_3 \\ 0 \\ \vdots \\ x_{N-1} \\ 0 \end{bmatrix} \tag{11}$$

At the position of zeros the root values are inserted as shown below

$$F^{-1}(x(n))\,|_{x(n)=x\left(n+\frac{N}{2}\right)} = \begin{bmatrix} x_1 \\ k_1 \\ x_3 \\ k_2 \\ \vdots \\ x_{N-1} \\ k_{\frac{N}{2}} \end{bmatrix} \tag{12}$$

These root value which are denoted by k are estimated in such a way that the value when propagated through the channel becomes zeros as shown below.

$$\begin{bmatrix} h_1^{ij} & \dots & h_{L_c}^{ij} & 0 & \dots & \dots & \dots & 0 \\ 0 & h_1^{ij} & \dots & h_{L_c}^{ij} & 0 & \dots & \dots & 0 \\ \vdots & & \ddots & & \ddots & & & \vdots \\ \vdots & \dots & 0 & h_1^{ij} & \dots & h_{L_c}^{ij} & 0 & 0 \\ \vdots & & & \ddots & \ddots & & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & h_1^{ij} & \dots & h_{L_c}^{ij} \end{bmatrix} \times \begin{bmatrix} x_1 \\ k_1 \\ x_2 \\ k_2 \\ \vdots \\ x_{\frac{N}{2}} \\ k_{\frac{N}{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{13}$$

The values are thus can be estimated as follows

$$\begin{bmatrix} h_2^{ij}k_1 & \dots & h_{L_c}^{ij}k_{\frac{L_c}{2}} & 0 & \dots & \dots & \dots & 0 \\ 0 & h_1^{ij}k_1 & \dots & h_{L_c}^{ij}k_{\frac{L}{2}} & 0 & \dots & \dots & 0 \\ \vdots & & \ddots & & \ddots & & & \vdots \\ \vdots & \dots & 0 & h_1^{ij}k_m & \dots & h_{L_c}^{ij}k_{m+\frac{L}{2}} & 0 & 0 \\ \vdots & & & \ddots & \ddots & & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & h_1^{ij}k_{\frac{L}{2}+1} & \dots & h_{L_c}^{ij}k_{\frac{N}{2}} \end{bmatrix} = \begin{bmatrix} -(h_1^{ij}x_1+\cdots+h_{L-1}^{ij}x_{L-1} \\ . \\ . \\ . \\ . \\ . \\ -(h_2^{ij}x_{\frac{L}{2}-1}+\cdots+h_{L-1}^{ij}x_{\frac{N}{2}} \end{bmatrix} \tag{14}$$

In the above equation the values $h_1^{ij}, h_2^{ij}, \dots h_{L_c}^{ij}$ and $x_1, x_2, \dots x_{\frac{N}{2}}$ are known values. Hence the value of $k_1, k_2, \dots k_{\frac{N}{2}}$ can be estimated simply by solving the equation.

6 BER Performance

This section analyses the performance of the Primary OFDM system when precoded with that of the conventional OFDM and Primary system without precoding. First the performance of the system is considered for an AWGN channel. The BER is plotted against the signal to noise ratio of the channel. The BER curve is shown in figure 6 for the proposed system. As shown in the figure the theoretical interference free system is found to have low BER. When two systems are used as cognitive users without precoding, then the BER performance of the primary system is not satisfactory. Even though the proposed system has BER little on the higher side, when compared with that of the non precoded system the BER is lower. The system performance is analyzed without using any scheme of encoding in order to determine the worst case performance. The constellation mapping used is 16bit QAM technique.

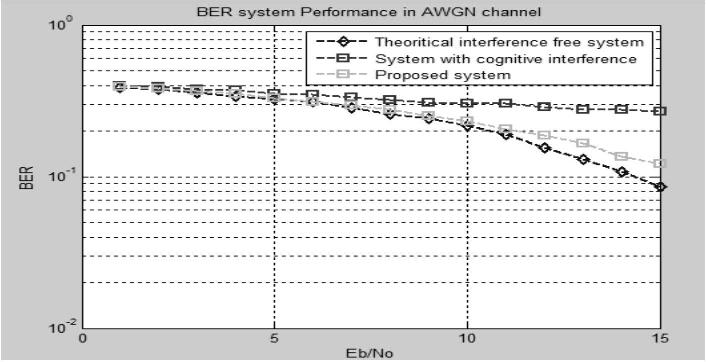


Fig. 6. BER performance in AWGN channel

As in any communication system not only the noise is the major phenomenon that impacts on BER, the ISI due to multipath propagation also poses a serious problem. Hence it is important to analyze the performance of the system in multipath channel. Hence the Rayleigh and Rician channel are considered for analysis. The figure below shows the performance of the system in Rayleigh channel. The BER is plotted against the Doppler shift measured in Hertz (Hz). The same system parameters are set as in case of AWGN channel. The cognitive interference is found to exhibit the maximum BER The proposed system is found to have reduced the BER at the primary due to precoding. The theoretical performance is having lower BER than the proposed system because. In the proposed system the precoding which is carried out is only done based on the channel state information available at the secondary receiver. This however cannot be stable all the time due to the random nature of noise which is added by the channel.

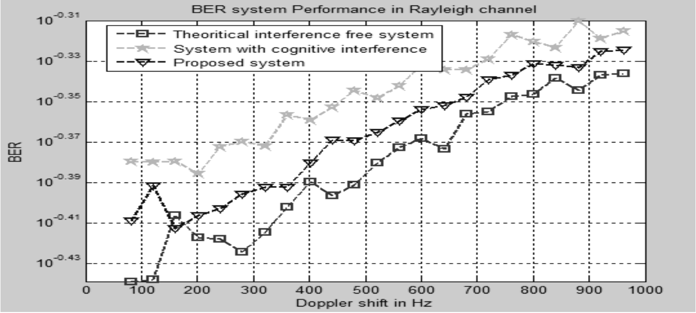


Fig. 7. BER performance in Rayleigh channel

Next the system is evaluated for performance in a Rician modeled channel. The Rician model is similar to that of the Rayleigh channel except for the fact that the Rician channel has direct line of sight transmission along with other line of transmission. The BER performance of the systems in Rician channel is shown in the figure below for varying Doppler shift. This motivates to estimate the performance of the system in Rician channel for varying direct path gains. This evaluation is depicted in the figure below. The system has a high BER value when the direct path gain is lower. This value tends to diminish as the value of direct path gains seems to increase.

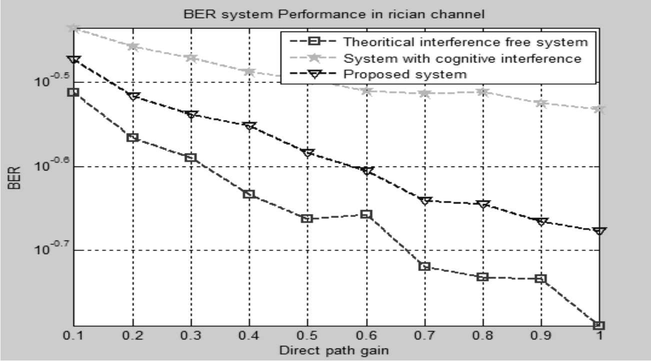


Fig. 8. BER performance in Rician channel for varying direct path gains

7 Conclusion

The cognitive radio is a technology which proves to be vital with the developing wireless systems. The precoded system is designed in order to reduce the cognitive interference caused by the secondary user to the primary user by insertion of root values. Which makes the recovery of messages at the secondary receiver much easier without need to have knowledge about the channel for which the transmitter is precoding as in the case of existing VFDM method. The graphical results obtained corroborates that the precoded OFDM can be used to reduce the BER caused by the cognitive interference. Thus the system can be used to eliminate the interference barring the fact that the secondary transmitter must know the channel state information precisely. The future works includes the analysis of different encoding scheme for the system. Also the consideration for hardware realization of the system is to be made.

References

1. Kolodzy, P., et al.: Next Generation Communications: Kickoff meeting. In: Proc. DARPA (October 17, 2001)
2. Devroye, N., Mitran, P., Tarokh, V.: Achievable rates in cognitive radio channels. IEEE Trans. on Inform. Theory 52(5), 1813–1827 (2006)
3. Molisch, A.F.: Wireless Communications, 2nd edn. Wiley (2011)
4. Dutta, A., Saha, D., Grunwald, D., Sicker, D.: Practical implementation of Blind Synchronization in NC-OFDM based Cognitive Radio Networks. In: CoRoNet 2010, Chicago, Illinois, USA (September 20, 2010)
5. Cardoso, L.S., Kobhayashi, M., Debbah, M.: Vondermonde Frequency Division Multiplexing for Cognitive Radio. arXiv:0803.0875v1 [cs. IT] (March 2008)

6. Cardoso, L.S., Cavalcanti, F.R.P., Kobayashi, M., Debbah, M.: Vondermonde-subspace Frequency Division Multiplexing receiver analysis. In: PIMRC, 2010 IEEE 21 International Symposium on Istanbul, Turkey (2010)
7. Maso, M., Cardoso, L.S., Bastug, E., Linh-Trung, N., Debbah, M., Ozdemir, O.: On the Practical Implementation of VFDM Based Oppurtunistic Systems: Issues and Challenges. REV Journal on Electronics and Communication 2(1-2) (January-June 2012)
8. Bokharaiee, S., Nguyen, H.H., Shwedyk, E.: Blind Spectrum Sensing for OFDM –Based Cognitive Radio System. IEEE Transaction on Vehicular Technology 60(3) (March 2011)
9. Ahmed, I., Arslan, T.: Design of multi standard turbo decoder for 3G and beyond, 1-4244-0630-7/07. IEEE (2007)
10. Zou, D., Lu, X., Xu, J., Wu, H.: Application of IFFT Based on CORDIC Algorithms in OFDM-UWB System. IEEE (2008)
11. Badoi, C.-I., Prasad, N., Croitoru, V., Prasad, R.: 5G based on Cognitive Radio. Wireless Pers. Communication (2011)
12. Mitola III, J.: Cognitive Radio for flexible mobile multimedia Communication. Moblie Networks and Application 6, 435–441 (2001)
13. Cabric, D., Mishra, S.M., Brodersen, R.W.: Implementation issues in spectrum sensing for Cognitive radio

Key Learning Features as Means for Terrain Classification

Ionut Gheorghe[1], Weidong Li[1], Thomas Popham[2], Anna Gaszczak[2],
and Keith J. Burnham[1]

[1] Control Theory and Applications Centre, Coventry University, Coventry CV1 5FB, UK
gheorghi@uni.coventry.ac.uk,{aa3719,ctac}@coventry.ac.uk
[2] Jaguar & Land Rover Research, University of Warwick, Coventry CV4 7AL, UK
{tpopham,agaszcza}@jaguarlandrover.com

Abstract. Modern vehicles seek autonomous subsystems adaptability to ever-changing terrain types in pursuit of enhanced drivability and maneuverability. The impact of key features on the classification accuracy of terrain types using a colour camera is investigated. A handpicked combination of texture and colour as well as a simple unsupervised feature representation is proposed. Although the results are restricted to only four classes {grass, tarmac, dirt, gravel} the learned features can be tailored to suit more classes as well as different scenarios altogether. The novel aspect stems from the feature representation itself as a global gist for three quantities of interest within each image: background, foreground and noise. In addition to that, the frequency affinity of the Gabor wavelet gist component to perspective images is mitigated by inverse homography mapping. The emphasis is thus on feature selection in an unsupervised manner and a framework for integrating learned features with standard off the shelf machine learning algorithms is provided. Starting with a colour hue and saturation histogram as fundamental building block, more complex features such as GLCM, k-means and GMM quantities are gradually added to observe their integrated effect on class prediction for three parallel regions of interest. The terrain classification problem is tackled with promising results using a forward facing camera.

Keywords: Terrain classification, machine learning, gist, GLCM, texture, colour, homography.

1 Introduction

Sensing is a critical aspect of vehicle operation, drivability and safety [9] and despite recent developments of radar, laser, and T-O-F (time of flight) [3], [14], the video camera still remains a desirable sensor for usage in ADAS (advanced driver assistance systems). It is a low cost sensing solution particularly suited for obstacle detection and scene interpretation. With drivability in mind, modern vehicles seek to self-adapt to driving conditions, traffic conditions as well as environment changes in terms of road surface. In particular, modern off-road capable vehicles subsystems are configurable depending on the terrain type. For example, different terrain types introduce different throttle response, suspension stiffness, locking differential etc.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_26, © Springer International Publishing Switzerland 2014

Currently, there is always a certain amount of reckoning from the driver that is needed for these changes to take place. This brings about the need for a sensing solution that would either assist the driver in taking the decision or help towards autonomous decision making altogether. In this work the focus is on terrain type prediction given colour images form a driving perspective. Four classes with potentially many different members but with similar intra class subsystems configurable properties are considered, namely {grass, tarmac, dirt, gravel}.

In previous literature, the terrain classification problem using colour camera [1], colour stereo camera and laser [12], [16] has been addressed mostly by the robotics community.

To date, the literature describing work on colour camera terrain/road change recognition is scarce [4], [8], [12], and [20]. It is this lack of investigation that motivates the decision to concentrate on what is suitable in terms of feature representation. It is necessary to find informative quantities for the terrain types in the form of feature combinations that capture underlying terrain/road properties within visual images. For that to come about, the feature representation is extrapolated from the more general concept of texture synthesis and image recognition using colour and texture. One of the daunting problems in the terrain domain is the intra class variety of visual words whereby it is hard to define spatial image prescriptions for recognition. A global formulation of ("gist" [19]) visual vocabulary is introduced, where spatial order is not important but quantitative order is, or equivalently the amount of data pooling corresponding to each word.

Without the latter observation this approach would stem entirely from classical B-o-F ("bag of features") restricted to allow for only few visual words for each class (i.e. background, foreground and spurious data). This granularity level of feature representation does not necessarily require word prescriptions for training and testing which means that the actual words can be used instead. Informally, in this representation of image saliency a choice is deliberately made to ignore spatial ordering of the "visual words". Their "significance" (contextual order) and "assertiveness" (data spread) provide valuable cues with respect to the terrain surface distribution. Normally, the contextual order can be extracted from a codebook in form of a word count but queries about the degree of feature "assertiveness" cannot be made. Since the intention is not to have large codebook with a rich vocabulary but a rather concise set of visual words to describe each image, the approach is taken to use the actual words for training and testing. Moreover, computational speedup is achieved by eliminating the codebook generation step as well as subsequent histogram features ("prescriptions") for classification [1]. The impact of using these visual words as part of the feature vector versus a standard texture descriptor (i.e. GLCM) is analyzed. This approach allows for multi-resolution classification but the processing time will scale accordingly as more data clogs the feature extraction pipeline.

2 Terrain Image Regions of Interest

Three ROIs are used for testing on a video sequence: left, middle and right (Figure 1). For classification based on colour and GLCM only, the image samples do not undergo geometric transformations. However, because the same texture appears to be at higher frequencies when samples are taken towards the vanishing point, feature pooling would produce different visual words.

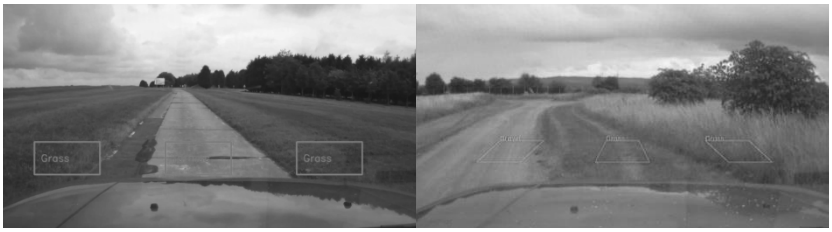


Fig. 1. Forward driving colour images

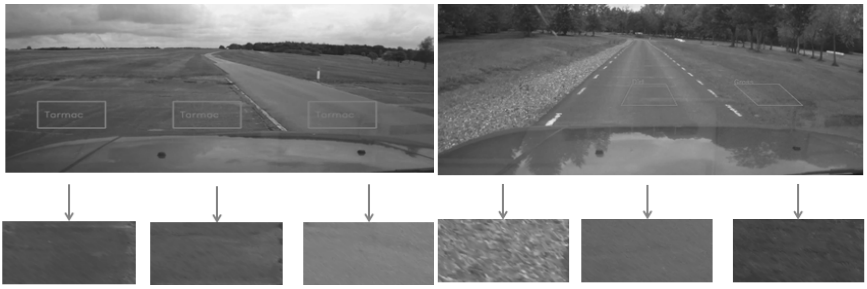


Fig. 2. Three samples taken from each test image

2.1 Rectangular Regions

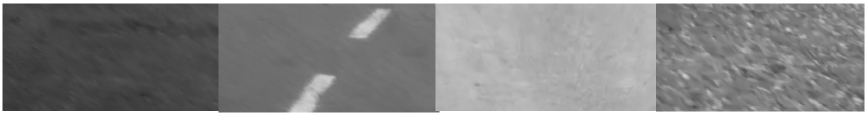
For colour and the non-wavelet texture descriptor (e.g. GLCM), the feature extraction is applied on rectangular ROIs with no homography rectification as it is shown in Figure 2 (left). Class prediction is based on pixel information of Hue and Saturation space and grey level co-occurrence measures.

2.2 Homography Rectified Regions

When the feature components include wavelets (i.e. Gabor wavelets), they are extracted from ROIs that undergo geometric transformations such as perspective transformations. In this way, an attempt to quantify texture in a robust way is made. By mitigating the impact of perspective image recording from forward facing camera at the road level, we try to preserve the true underlying road texture. First, we extract images from two lateral parallelograms (accounting for horizontal shear) and one middle trapezium as in Figure 2 (right). Then these images get warped into rectangular images of 200 x 100 pixels and feature extraction follows. The homography matrix describes how the warping takes place form the original ROIs to the new rectangular images. It is possible to find homography automatically, however previous attempts rely on key point matching between two images and are slow in practice. For example [17] uses SIFT and RANSAC to estimate homography. Experiments have been carried out with just a handpicked road level transformation that maps predefined regions to rectangles in a fixed amount of time. Furthermore, these regions are manually chosen to cover the road view as a trapezium and the road sides as parallelograms for a forward driving perspective. Figure 2 (right) shows how this rectification takes place. It also shows one of the daunting problems present in our dataset, namely the bonnet reflection onto the windshield which creates colour artifacts and occasionally results in erroneous classification.

3 Feature Extraction and Representation

Typical training examples for {grass, tarmac, dirt, gravel}



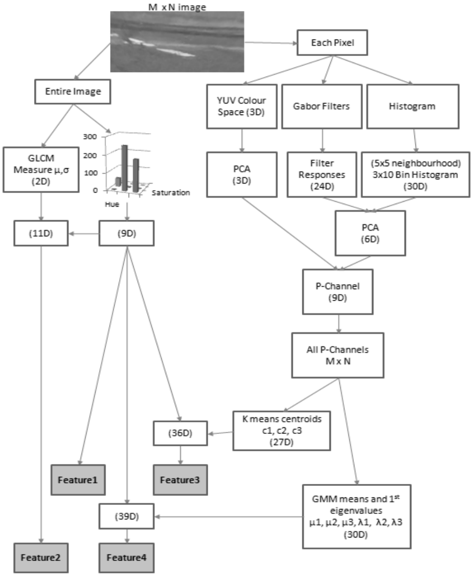


Fig. 3. Concatenating dimensions to form feature vectors

Four categories of features are considered for training and testing the terrain classes (Figure 3). The first two are handpicked features and the last two require unsupervised learning applied to p-channels [10]:

1. Hue and saturation histogram – a 9D vector is formed by concatenating the lines of a 2 dimensional histogram containing the 3 Hue Bins x 3 Saturation Bins normalized to 255. This becomes the starting point for other more complex feature representations. Mere colour information is critical at the most basic representation level. The 9D Bins vector is kept fixed although its granularity can be changed.

2. The 9D HS Bins concatenated with a 2D GLCM measure for 4 directions (μ, σ). The 2D GLCM representations include: entropy, homogeneity, energy, maximum probability and contrast.
3. The 9D HS Bins concatenated with the k-means (3 clusters) centroids c1, c2, c3 in the following order: background, foreground and noise.
4. The 9D HS Bins concatenated with the GMM means (3 Gaussians) μ1, μ2, μ3, and largest covariance eigenvalues for each Gaussian λ1, λ2, λ3 in the following order: background, foreground and noise.

4 Method Discussion

Performance is progressively tested by introducing increasingly complex features starting with colour, GLCM [20], and finishing off with unsupervised feature selection. Illumination changes in the outdoor environment are implicitly accounted for by discarding the luminance (or value) component from our HSV space [15] and to some extent, by using a filter bank of 24 Gabor filters (4 scales and 6 orientations) as part of our texture descriptor. This is also done explicitly by means of training examples representative for the intra class variance. Robustness is achieved using average pooling on normalized p-channels (in the form of cluster centroids or GMM means) representing both the overall texture and colour. When fitting a GMM, we restrict the covariance of each Gaussian to be either spherical or diagonal. Moreover, to introduce (i.e. quantify) spread information but avoid directionality, only the (largest) eigenvalues of GMM covariance matrixes are added to the feature vectors and their corresponding eigenvectors are ignored. Given the previous assumptions this becomes equivalent to using the variance, for spherical covariance GMM or the maximum dimensional variance for diagonal covariance GMM. The purpose of the latter is to quantify the largest data spread occurring in one of the data dimensions. Normally, if directionality was needed, that alone would not have sufficient discriminative power due to the fact that it is plausible to have similar examples whereby the largest data spread would occur in different dimensions. In fact, the covariance eigenvectors can be viewed as versor directions and the eigenvalues act as scaling quantitative descriptors for the p-channels spread in a Hilbert space.

4.1 HS Bins

The HS Bins 9D vector is the main colour cue used for composing the feature vector. It is extracted from a 3x3 Histogram and its values are normalized to 255 (e.g. if all the pixels belong to a single bin that dimension would store 255 and all other 0).

4.2 GLCM Statistics

Grey level co-occurrence matrix quantities are computed for four directions in each image at E, N-E, N, and N-W. Figure 4 shows contrast (left) and maximum probability (right) plotted as mean vs. standard deviation on all directions for the entire training set.

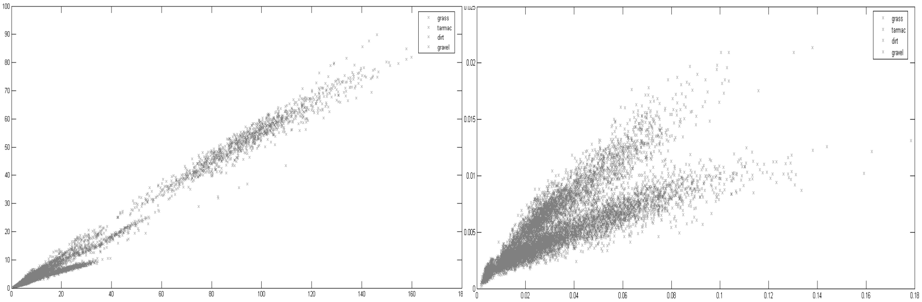


Fig. 4. 2D clusters from GLCM representing each terrain type

4.3 K-means Applied to p-Channels

It has been recently proven that the cluster centroids subspaces are the same as PCA subspaces represented by principal components [7]. Thus projecting raw data into PCA space can be used as e bootstrap procedure for k-means. PCA enables k-means to run into a space where the global optimum cluster configuration lies. In fact, a recent study [6] showed better performance of unsupervised feature learning algorithms such as k-means, GMMs on whitened data as opposed to raw inputs. The method suggested in [10] is used for dimensionality reduction. Principal channels are extracted as a combination of colour and texture for the equivalent number of pixels within each image ROI and then run k-means.

Generating Vocabulary Using k-Means

A vocabulary is generated using k-means on the p-channels by taking the cluster centroids as words. Instead of indexing the codebook (1-of-k words) containing the words as in previous approaches [13], the actual words are used for training in a certain semantic order. A form of average pooling is applied in the sense that the cluster centroids locations are at the local average location of the p channels. Other forms of data pooling such as maximum pooling [13] have also been plausible but for this definition of terrain foreground, background and noisy quantities averaging suffices.

K-means Learning Centers

The k-means cluster centroids are extracted as learned centers (or features) and ranked according to the number of samples contained in each cluster. An improved version of the k-means algorithm is used to deal with the cluster initialization in a probabilistic manner [2]. The first cluster center is drawn randomly from a uniform distribution and subsequent centers are drawn randomly from a multivariate distribution whereby they are more likely to fall close to their actual location. Clustering is initialized and ran three times and only the most compact configuration is retained.

$$\sum_i ||samples_i - centers_{labels_i}||^2 \tag{1}$$

Although the k-means algorithm is not required to converge to the exactly same cluster configuration, a distance measure is defined on the data set to prove that the cluster centroids are well suited features for learning. Centroids and the distance between them are computed for all training examples to find the smallest Euclidean norm between two possible cluster centers. This is equivalent to finding the most similar words in the entire vocabulary in terms of semantics. This norm is then used as a reference and k-means is run multiple times for uniform images in both texture and colour. Finding a stable multi-cluster configuration for rather uniform images is a difficult task but k-means produced the same clustering almost every time. The most inconsistent cluster configuration was found to be just within ~1 % of the smallest Euclidean norm on the entire training set, thus making centroids suitable features to learn. Intuitively, this test ensures that the semantics of certain words in the visual vocabulary remain unchanged from a contextual point of view. For the classification problem it was sufficient to initialize three clusters to account for background, foreground and spurious quantities respectively, within the training and testing images.

4.4 GMM Means and Data Spread Information on P-Channels

While the previous clustering technique learns very informative tendencies about the data structure, it lacks the much needed information about the spread (Figure 5). In terms of semantic interpretation this is the ambiguity of the vocabulary words. To overcome this problem, the previously obtained cluster centers are fed as means for the GMM [5] and the distribution parameters for three Gaussians are learned. The EM algorithm learns parameters such as the covariance matrixes from the data spread for each distribution. Unless a specific kind of Gaussian is imposed on the algorithm, the means remain similar to k-means. At this stage the feature vector can be enlarged to contain not only the Gaussian means but also their corresponding most significant covariance eigenvalues and eigenvectors.

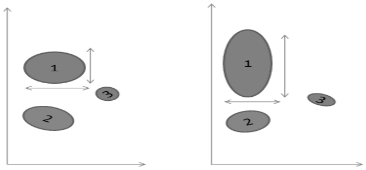


Fig. 5. In this illustrative example k-means generates same centroids in both cases

5 Machine Learning Applied on Features

The classification approach follows that of DAGSVM presented in [18]. In [11], the practicality of DAGSVM and one-vs-one as solutions to SVM multiclass labeling is argued. However for this four class problem a decision was taken to capitalize only on DAGSVM where each node makes prediction using a linear kernel (Figure 6). Training was done on 10000 images containing the four classes {grass, tarmac, dirt, gravel} under different illumination conditions and mild weather changes with samples taken from videos shot in different days. Testing was carried out on a separate video file with a length of 7000 frames not seen during the training process.

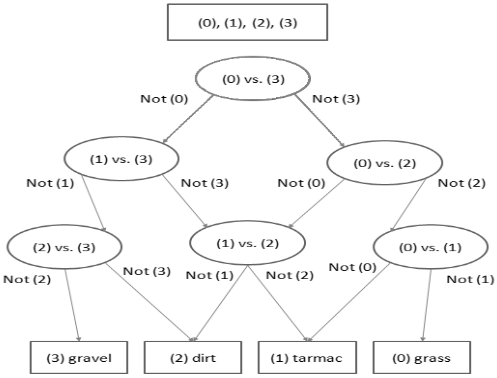


Fig. 6. DAGSVM requires $C^2_{Classes}$ nodes

6 Experimental Results

Experiments have been run on an Intel Core I7-2670QM CPU @ 2.20 GHz clock rate. The same algorithm running on such machine would vastly benefit from allowing for various parts of the algorithm to be executed by parallel threads on multiple cores. The average performance and classification time for the three regions of interest are summarized in Table 1.

Table 1. Performace of used features

Feature representation	Classification (%)	Time (ms)
HS bins	46.19	1.32
HS bins + GLCM Entropy μ,σ	49.71	26.71
HS bins + GLCM Homogeneity μ,σ	50.64	24.36
HS bins + GLCM Energy μ,σ	53.04	25.98
HS bins + GLCM Maximum Probability μ,σ	61.12	24.92
HS bins + GLCM Contrast μ,σ	64.32	26.59
HS bins + k-means c1, c2, c3	85.92	3861.37
HS bins + GMM means μ1, μ2, μ3+eigenvalues λ1, λ2, λ3	94.75	6242.28

Best confusion matrix (Table 2) shows an accuracy of 95.08 % using the 39D feature vector made up of HS bins + GMM means μ1, μ2, μ3 + eigenvalues λ1, λ2, λ3 with an average prediction time of 5.7 s.

Table 2. Confusion matrix showing the hit rate

		Predicted class			
		grass	tarmac	dirt	gravel
True class	grass	0.95	0.01	0.02	0.02
	tarmac	0.01	0.94	0.04	0.01
	dirt	0.03	0.02	0.94	0.01
	gravel	0.01	0.01	0.01	0.97

7 Conclusion

A novel feature representation has been proposed as a combination of handpicked features and single layered learned features to tackle terrain recognition with good accuracy. Investigation started with computationally cheap features and more complex features were added for better classification performance. Hue and saturation colour space (value aside) contributed with histogram bins as dimensions to all features in the attempt to produce a colour bias and achieve robustness to mild illumination changes. Moreover, perspective warping was applied before using wavelet texture descriptors. This is to prevent frequency responses from contributing to wrong clustering in the Hilbert space gist formulation. There is a clear trade-off between representation and classification time and method optimisation will follow.

References

1. Angelova, A., Matthies, L., Helmick, D.M., Perona, P.: Fast Terrain Classification Using Variable-Length Representation for Autonomous Navigation. In: CVPR (2007)
2. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, p. 1027. Society for Industrial and Applied Mathematics, Philadelphia (2007)
3. Badino, H., Huber, D., Kanade, T.: Integrating LIDAR into Stereo for Fast and Improved Disparity Computation. In: 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), p. 405 (2011)
4. Beucher, S., Yu, X.: Road recognition in complex traffic situations. In: Proc. 7th IFAC/IFORS Symp. Transp. Syst.: Theory Appl. Adv. Technol., pp. 413–418 (1994)
5. Bishop, C.M.: Pattern recognition and machine learning. Springer, New York (2006)
6. Coates, A., Ng, A.Y., Lee, H.: An Analysis of Single-Layer Networks in Unsupervised Feature Learning. Journal of Machine Learning Research - Proceedings Track 15, 215–223 (2011)
7. Ding, C., He, X.: K-means clustering via principal component analysis. In: ICML 2004: Proceedings of the Twenty-first International Conference on Machine Learning, p. 29. ACM, New York (2004)
8. Fernandez-Maloigne, C., Bonnet, W.: Texture and neural network for road segmentation. In: Proc. Intell. Veh. Symp., pp. 344–349 (1995)
9. Fleming, W.J.: New Automotive Sensors—A Review. IEEE Sensors Journal 8(11), 1900–1921 (2008)

10. Gavish, L., Shapira, L., Wolf, L., Cohen-Or, D.: One-sided object cutout using principal-channels. In: 11th IEEE International Conference on Computer-Aided Design and Computer Graphics, CAD/Graphics 2009, p. 448 (2009)
11. Hsu, C., Lin, C.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13(2), 415–425 (2002)
12. Jansen, P., van der Mark, W., van den Heuvel, J.C., Groen, F.C.A.: Colour based off-road environment and terrain type classification. In: Proceedings 2005 IEEE Intelligent Transportation Systems, p. 216 (2005)
13. Jianchao, Y., Kai, Y., Yihong, G., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, p. 1794 (2009)
14. Jiejie, Z., Liang, W., Ruigang, Y., Davis, J.: Fusion of time-of-flight depth and stereo for high accuracy depth maps. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, p. 1 (2008)
15. Lefèvre, S., Vincent, N.: Efficient and robust shot change detection. J. Real-Time Image Processing 2(1), 23–34 (2007)
16. Manduchi, R., Castano, A., Talukder, A., Matthies, L.: Obstacle detection and terrain classification for autonomous off-road navigation. Autonomous Robots 18, 81–102 (2005)
17. Moisan, L., Moulon, P., Monasse, P.: Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers. Image Processing On Line (2012)
18. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large Margin DAGs for Multiclass Classification. In: Advances in Neural Information Processing Systems, p. 547. MIT Press (2000)
19. Siagian, C., Itti, L.: Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(2), 300–312 (2007)
20. Tang, I., Breckon, T.P.: Automatic Road Environment Classification. IEEE Transactions on Intelligent Transportation Systems 12(2), 476–484 (2011)

Manifold Regularized Particle Filter for Articulated Human Motion Tracking

Adam Gonczarek and Jakub M. Tomczak

Institute of Computer Science, Wrocław University of Technology,
wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{adam.gonczarek,jakub.tomczak}@pwr.wroc.pl

Abstract. In this paper, a fully Bayesian approach to articulated human motion tracking from video sequences is presented. First, a filtering procedure with a low-dimensional manifold is derived. Next, we propose a general framework for approximating this filtering procedure based on the particle filter technique. The low-dimensional manifold can be treated as a regularizer which restricts the space of all possible distributions to the space of distributions concentrated around the manifold. We refer to our method as *Manifold Regularized Particle Filter*. The proposed approach is evaluated using real-life benchmark dataset *HumanEva*.

Keywords: articulated motion tracking, manifold regularization, particle filter, generative approach, Gaussian process latent variable model.

1 Introduction

Articulated human motion tracking from video image sequences is one of the most challenging computer vision problems for the past two decades. The basic idea behind this issue is to recover a motion of the complete human body basing on the image evidence from a single or many cameras, and without using any additional devices, e.g., color or electromagnetic markers. Such system can be applied in many everyday life areas, see [8,9]. Pointing out only a few, motion tracking may be used in control devices for Human-Computer Interaction, surveillance systems detecting unusual behaviors, dancing or martial arts training assistants, support systems for medical diagnosis.

During last years, a lot of effort has been put in solving the human motion tracking issue. However, excluding some minor cases, the problem is still open. There are several reasons worth mentioning that make the issue very difficult. First, there is a huge variety of different images corresponding to the same pose that may be obtained. This is caused by variability in human wear and appearance, changes in lighting conditions, camera noise, etc. Second, image lacks of depth information which makes impossible to obtain three-dimensional pose from two-dimensional images. Moreover, one has to handle different types of occlusions including self-occlusions and occlusions caused by external environment. Finally, efficient exploration of the space of all possible human poses is troublesome because of high-dimensionality of the space and its non-trivial constraints.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_27, © Springer International Publishing Switzerland 2014

To date, however, several conceptually different approaches have been proposed to address the human motion tracking problem. They can be roughly divided into two groups. In the first one, discriminative methods are used to model directly the probability distribution over poses conditioned on the image evidence, see [1,3,7]. On the contrary, in the second group a generative approach is used to model separately the prior distribution over poses and the likelihood of how well a given pose fits to the current image. Pure generative modeling assumes that one tries to model the true pose space as accurately as it is possible and uses Bayesian inference to estimate current pose, see [2,6,11,13,14]. Recent studies show that using more flexible models, i.e., part-based models and searching maximum a posteriori estimate (MAP) give very promising results, see [15]. However, these approaches are mainly applied to 2D pose estimation problems.

The contribution of the paper is threefold:

1. A general framework for human motion tracking using generative modeling and hidden low-dimensional manifold is proposed.
2. A particle filter regularized using low-dimensional manifold is introduced. Further, we refer to this particle filter as *Manifold Regularized Particle Filter* (MRPF).
3. A dynamics model using Gaussian process latent variable model (GPLVM) and low-dimensional manifold is presented.

Empirical results are evaluated on benchmark dataset HumanEva, see [11] for details.

The paper is organized as follows. In Section 2 the problem of the human motion tracking is outlined. First, the problem of pose estimation is given in section 2.1 and then the human motion tracking problem is stated in section 2.2. Next, the likelihood function is formulated in section 4. In Section 3 the particle filter with low-dimensional manifold is proposed. The model of dynamics with low-dimensional manifold is presented in section 5. At the end, the empirical study is carried out in section 6 and conclusions are drawn in section 7.

2 Human Motion Tracking

In this paper, we assume that a human body is represented by a set of articulately connected rigid parts. Each connection between two neighboring elements characterize a joint and can be described by up to three degrees of freedom, depending on movability of the joint. All connected parts form a kinematic tree with the root typically associated with the pelvis. A common representation of the state of the k^{th} joint uses Euler angles which describe relative rotation between neighboring parts in the kinematic tree. However, we prefer to use quaternions because they can be compared using the Euclidean metric. In case of angles of rotation of the parts in the kinematic tree in the range between 0 and π we can take advantage of an approximation of quaternions (for details see [12]).

The set of quaternions for all K joints together with the global position and orientation of the kinematic tree in 3D constitutes the minimal set of variables that are used to describe the current state of the human body, which is denoted by $\mathbf{x}$. It is worth mentioning that $\mathbf{x}$ is usually around 40-50 dimensions, which is one of the fundamental reasons that makes the human motion tracking a difficult problem.

We assume that there are several synchronized cameras which provides video images of a human body from different perspectives. The cameras should be located so that to contribute as much information about the body as possible, i.e., they should register different parts of a scene. Let $\mathcal{I}$ denote a set of all available images from all cameras at current moment. Hence, we want to infer the human body configuration $\mathbf{x}$ basing on $\mathcal{I}$.

In next sections, first we formulate pose estimation problem and then state the human motion tracking problem.

2.1 Pose Estimation Problem Statement

The goal of the pose estimation issue is to find a human pose estimate $\hat{\mathbf{x}}$ in every video frame, basing on all available images $\mathcal{I}$. Since this is a typical multivariate regression problem and the optimal solution (in the sense of decision making), that is, $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x}|\mathcal{I}]$.

The key issue is to model properly the true distribution $p(\mathbf{x}|\mathcal{I})$. It can be accomplished using discriminative models, i.e., explicit modeling of $p(\mathbf{x}|\mathcal{I})$, or generative models by applying Bayes' rule, $p(\mathbf{x}|\mathcal{I}) \propto p(\mathcal{I}|\mathbf{x})p(\mathbf{x})$, and then modeling $p(\mathcal{I}|\mathbf{x})$ and $p(\mathbf{x})$ separately. In this paper, we focus on the second approach.

2.2 Human Motion Tracking Problem Statement

Now let us extend the problem of the pose estimation to tracking the whole sequence of states $\mathbf{x}_{1:T} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, also called a *trajectory*. Additionally, we need to maintain the sequence of images $\mathcal{I}_{1:T} = \{\mathcal{I}_1, \dots, \mathcal{I}_T\}$. The sequence of video frames corresponds to images achieved from all cameras in consecutive moments $1, \dots, T$.

It is a fact that the high-dimensional pose space consists of human body configurations and most of them are unrealistic. Additionally, during specific motions (e.g. walking or running) all state variables exhibit strong correlations which depends on the current pose. These two remarks yields a corollary that the real trajectories of motion form a low-dimensional manifold.

We assume that any state $\mathbf{x}$ corresponds to a point on a low-dimensional manifold $\mathbf{z}$. Hence, the trajectories of human motion formulate a pattern which locally oscillates around the low-dimensional manifold. Further, we are interested in representing the joint probability distribution $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathcal{I}_{1:T})$. The manner how it is factorized is graphically presented by the probabilistic graphical model in Figure 1. Notice that the current state $\mathbf{x}_t$ influences future state and future point on the manifold $\mathbf{z}_{t+1}$ which in turn impacts $\mathbf{x}_{t+1}$. In the literature, several similar models have been proposed, e.g., in [16] a model assumes that the temporal dependence exists between low-dimensional variables only, in

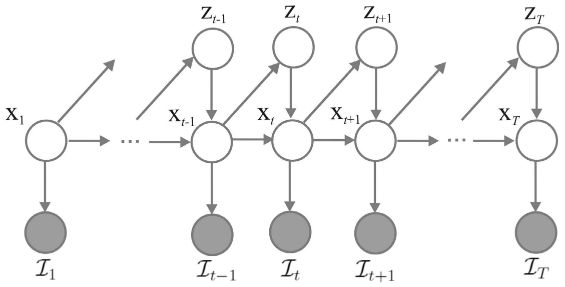


Fig. 1. Probabilistic graphical model with a low-dimensional manifold represented by variables $\mathbf{z}$

[13] a conditional Restricted Boltzmann Machine is used to represent information about the low-dimensional manifold which leads to undirected dependence between $\mathbf{x}$ and $\mathbf{z}$.

We are interested in calculating the a posteriori probability distribution for $\mathbf{x}_t$ given images $\mathcal{I}_{1:t}$ by marginalizing $p(\mathbf{x}_{1:t}, \mathbf{z}_{1:t}|\mathcal{I}_{1:t})$ over all state variables $\mathbf{x}_{1:t-1}$ and hidden variables $\mathbf{z}_{1:t}$ which yields:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) = \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \iint p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}\mathrm{d}\mathbf{z}_t, \tag{1}$$

where $p(\mathcal{I}_t|\mathcal{I}_{1:t-1})$ is the normalization constant.

We have obtained a filtering procedure which includes information about the low-dimensional manifold. Further, we need to determine the following aspects:

- an algorithm which allows to perform the filtering procedure (1);
- the likelihood function $p(\mathcal{I}_t|\mathbf{x}_t)$;
- models of dynamics: $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ and $p(\mathbf{z}_t|\mathbf{x}_{t-1})$.

3 Manifold Regularized Particle Filter

In the context of the human motion tracking the filtering procedure is intractable and hence an approximation of (1) should be applied. Typically, a sampling method like a particle filter technique is applied. However, the main disadvantage of the particle filter is that it needs to generate a huge amount of particles in order to cover a high-dimensional state space. Otherwise, it fails to approximate true distribution. In order to cover the highly probable areas in the pose space only, an extension of the particle filter technique, called *Annealed Particle Filter* (APF), has been proposed [2]. However, this method tends to be trapped in one or a few dominating extrema. Therefore, in the context of the human motion tracking, it is non-robust to noisy likelihood model and thus fails to track the proper trajectory.

In this paper, we propose a different approach which modifies the particle filter by introducing a regularization in a form of the low-dimensional manifold. This filtering procedure operates in the neighborhood of the low-dimensional space where the true poses are concentrated, and thus it guarantees that highly probable regions are covered and the particles are distributed around different local extrema.

First, let us remind that we want to calculate $p(\mathbf{x}_t|\mathcal{I}_{1:t})$. If we were able to sample from $p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$, it would be possible to approximate (1), concentrated on points $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)} \sim p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \sum_{n=1}^{N} \pi(\mathbf{x}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{2}$$

where $\pi(\mathbf{x}_t^{(n)})$ is a normalized form of a single score calculated using the following expression $\tilde{\pi}(\mathbf{x}_t^{(n)}) = p(\mathcal{I}_t|\mathbf{x}_t^{(n)})$.

Usually, it is troublesome to generate $\mathbf{x}_t$ for given $\mathbf{x}_{t-1}$ and $\mathbf{z}_t$ using the distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ and thus we will introduce an auxiliary distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. Then, taking advantage of dependencies defined by the probabilistic graphical model in Figure 1, we get:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = \frac{1}{Z}\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)Q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}), \tag{3}$$

where $\tilde{\omega}$ are weights:

$$\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \tag{4}$$

and Q is an auxiliary joint distribution:

$$Q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}), \tag{5}$$

and Z is a normalization constant.

Eventually, we can approximate the a posteriori distribution (1) using the following formula:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{6}$$

where the normalized weights are defined as follows:

$$\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) = \frac{\tilde{\pi}(\mathbf{x}_t^{(n)})\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})}{\sum_{j=1}^{N} \tilde{\pi}(\mathbf{x}_t^{(j)})\tilde{\omega}(\mathbf{x}_t^{(j)}, \overline{\mathbf{x}}_{t-1}^{(j)}, \mathbf{z}_t^{(j)})}. \tag{7}$$

Notice that we introduce the low-dimensional manifold in the manner that the particles are weighted by ω. The procedure of MRPF is presented in Algorithm 1.

Algorithm 1. Manifold Regularized Particle Filter

```
Input  : initial state $\mathbf{x}_0$, sequence of images $\mathcal{I}_{1:T}$
Output: sequence of state estimates $\hat{\mathbf{x}}_{1:T}$
Duplicate the initial state $\mathbf{x}_0$ and formulate a set: $\overline{\mathcal{X}}_0 = \{\overline{\mathbf{x}}_0^{(1)}, \ldots, \overline{\mathbf{x}}_0^{(N)}\}$ ;
for $t = 1 : T$ do
    Generate a sample $\mathcal{Z}_t = \{\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(N)}\}$ using $\mathbf{z}_t^{(n)} \sim p(\mathbf{z}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
    Generate a sample $\mathcal{X}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)}\}$ using $\mathbf{x}_t^{(n)} \sim q(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
    Calculate $\tilde{\pi}(\mathbf{x}_t^{(n)})$ using the likelihood model $p(\mathcal{I}_t|\mathbf{x}_t)$ ;
    Calculate $\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ using (4);
    Normalize the weights $\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ using (7);
    Calculate the estimate of the state variables
    $\hat{\mathbf{x}}_t = \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) \mathbf{x}_t^{(n)}$ ;
    Generate a sample $\overline{\mathcal{X}}_t = \{\overline{\mathbf{x}}_t^{(1)}, \ldots, \overline{\mathbf{x}}_t^{(N)}\}$ using approximation (6);
end
```

4 Likelihood Function

The likelihood function $p(\mathcal{I}_t|\mathbf{x}_t)$ aims at evaluating the given human body configuration $\mathbf{x}_t$ corresponds to the set of images $\mathcal{I}_t$. We compare images which contain a human body model projected onto camera views with binary silhouettes obtained from background subtraction procedure, by calculating the difference between them. This model is called *bidirectional silhouette likelihood* [11].

5 Dynamics Model Using Low-Dimensional Manifold

The simplest model used for modeling the dynamics of the pose $\mathbf{x}$ is a Gaussian diffusion, i.e., new state is the old state disturbed by an independent Gaussian noise. However, this model turns to be too simplistic in the context of the human motion tracking. Therefore, we propose to model the dynamics of the pose using low-dimensional manifold and a nonlinear dependency. First, we need to learn the low-dimensional manifold. Second, a model for dynamics on the low-dimensional manifold has to be proposed, $p(\mathbf{z}_t|\mathbf{x}_{t-1})$. Third, the model of dynamics in the pose space with the low-dimensional manifold should be given, $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$.

5.1 Learning the Low-Dimensional Manifold

For learning the low-dimensional manifold we apply the *Gaussian Process Latent Variable Model* (GPLVM) [4]. The GPLVM model constitutes a non-linear dependency between the pose and the low-dimensional manifold as follows: $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}$, where i^{th} function is a realization of the Gaussian process [10], $f_i \sim \mathcal{GP}(f|0, k(\mathbf{z}, \mathbf{z}'))$, where k is a kernel function, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|0, \sigma_z^2\mathbf{I})$, where σ_z^2 is variance, and $\mathbf{I}$ denotes the identity matrix. In this paper, we use the RBF kernel, $k(\mathbf{z}, \mathbf{z}') = \beta \exp\left(-\frac{\gamma_z}{2}\|\mathbf{z} - \mathbf{z}'\|^2\right) + \beta_0$.

We are interested in finding a matrix of low-dimensional variables corresponding to observed poses, i.e., a matrix $\mathbf{Z}$ for observed poses $\mathbf{X}$. Additionally, we want to determine the mapping between the manifold and the high-dimensional space by learning parameters β, β_0 and γ_z, and σ_z^2. The training corresponds to finding the parameters and points on the manifold that maximize the logarithm of the likelihood function in the following form:

$$\begin{aligned}\ln p(\mathbf{X}|\mathbf{Z}) &= \ln \prod_{i=1}^{D} \mathcal{N}(\mathbf{X}_{:,i}|0, \mathbf{K} + \sigma_z^2 \mathbf{I}_{T\times T}) \\ &= -\frac{DT}{2}\ln(2\pi) - \frac{D}{2}\ln|\overline{\mathbf{K}}| - \frac{1}{2}\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\mathbf{X}), \qquad (8)\end{aligned}$$

where $\mathbf{X}_{:,i}$ denotes i^{th} column of the matrix $\mathbf{X}$, $|\cdot|$ and $\mathrm{tr}(\cdot)$ are matrix determinant and trace, respectively, $\overline{\mathbf{K}} = \mathbf{K} + \sigma_z^2\mathbf{I}_{T\times T}$, and $\mathbf{K} = [k_{nm}]$ is the kernel matrix with elements $k_{nm} = k(\mathbf{z}_n, \mathbf{z}_m)$.

Let us notice that solutions of the maximization $\mathbf{z}_t$ and γ_z can be arbitrarily re-scaled, thus there are many equivalent solutions. In order to avoid this issue we introduce a regularizer $\frac{1}{2}\|\mathbf{Z}\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm, and the final objective function takes the form $L(\mathbf{Z}) = \ln p(\mathbf{X}|\mathbf{Z}) - \frac{1}{2}\|\mathbf{Z}\|_F^2$. The objective function can be optimized using standard numerical optimization algorithms, e.g., scaled conjugate gradient method. Additionally, the objective function is not concave and hence have multiple local maxima. Therefore, it is important to initialize the numerical algorithm properly, e.g., by using principal component analysis.

The kernel function used to determine the covariance function takes high values for points $\mathbf{z}_n$ and $\mathbf{z}_m$ that are close to each other, i.e., they are similar. Moreover, because the points on the manifold are similar, the original poses $\mathbf{x}_n$ and $\mathbf{x}_m$ are similar as well. However, the situation does not hold in the opposite direction. This issue is undesirable in the proposed filtering procedure 1 because the distribution $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ is multi-modal and thus hard to determine. However, this effect can be decreased by introducing *back constraints* which leads to *Back-Constrained* GPLVM (BC-GPLVM) [5].

The idea behind BC-GPLVM is to define $\mathbf{z}$ as a smooth mapping of $\mathbf{x}$, $\mathbf{z} = \mathbf{g}(\mathbf{x})$. For example, this mapping can be given in the linear form, i.e., $g_i(\mathbf{x}) = \sum_{t=1}^{T} a_{ti}k_x(\mathbf{x}, \mathbf{x}_t) + b_i$, where g_i denotes i^{it} component of $\mathbf{z}$, a_{ti}, b_i are parameters, and $k_x(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\gamma_x}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right)$ is the kernel function in the high-dimensional space of poses. We can incorporate the mapping into the objective function, i.e., $z_n^i = g_i(\mathbf{x}_n)$, and then optimize w.r.t. a_{ti} and b_i instead of z_n^i. The application of back constrains entails closeness of low-dimensional points $\mathbf{z}_t$ if high-dimensional points $\mathbf{x}_t$ are similar.

The big advantage of Gaussian processes is tractability of calculating the predictive distribution for new pose $\mathbf{x}_p$ and its low-dimensional representation $\mathbf{z}_p$. The corresponding kernel matrix is as follows:

$$\begin{bmatrix} \overline{\mathbf{K}} & \overline{\mathbf{k}} \\ \overline{\mathbf{k}}^{\mathrm{T}} & \overline{k}_z(\mathbf{z}_p, \mathbf{z}_p) \end{bmatrix}, \qquad (9)$$

and finally the predictive distribution [10]:

$$p(\mathbf{x}_p|\mathbf{z}_p, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{x}_p|\boldsymbol{\mu}_p, \sigma_p^2 \mathbf{I}_{D\times D}), \tag{10}$$

where:

$$\boldsymbol{\mu}_p = \mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}, \tag{11}$$

$$\sigma_p^2 = \overline{k}_z(\mathbf{z}_p, \mathbf{z}_p) - \overline{\mathbf{k}}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}. \tag{12}$$

5.2 Dynamics on the Manifold

Idea of the model $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ is to predict new position on the manifold basing on the previous pose. Therefore, we need a mapping which allows to transform a high-dimensional representation to a low-dimensional one. For this purpose we apply the back constraints. Adding Gaussian noise with the covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}^2_{x\to z})$ to the back constraints, we obtain the following model of the dynamics on the manifold:

$$p(\mathbf{z}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{g}(\mathbf{x}_{t-1}), \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to z})). \tag{13}$$

5.3 Dynamics in the Pose Space with the Low-Dimensional Manifold

The model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ determines the probability of the current pose basing on the previous pose and the current point on the low-dimensional manifold. A reasonable assumption is that the model factorizes into two components, namely, one concerning only previous pose, and second – the low-dimensional manifold. This factorization follows from the fact that these two quantities belong to two different spaces and thus are hard to compare quantitatively. Then, the model of dynamics takes the following form:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) \propto p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t). \tag{14}$$

The first component is expressed as a normal distribution with the diagonal covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x})$:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x})). \tag{15}$$

The second component is constructed using the mean of the predictive distribution (11) and is disturbed by a Gaussian noise with the diagonal covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})$ which leads to the following model:

$$p(\mathbf{x}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})). \tag{16}$$

It is important to highlight that the training of the parameters $\mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})$ has to be performed using a separate validation set which contains data. Otherwise, using the same training set as for determining $\mathbf{Z}$ leads to underestimation of the parameters.

5.4 Dynamics Models and the Filtering Procedure

At the end, let us consider the application of the particle filter proposed earlier (see Algorithm 1) in the context of the outlined models of dynamics. First, we need to propose the auxiliary distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. In our case it is given in the form (15), i.e., $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x}))$. Then, the weights $\tilde{\omega}$ are given in the form (16), i.e., $\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x}))$.

6 Empirical Study

The aim of the experiment is to compare the proposed approach using MRPF with methods using the ordinary particle filter (PF) and the annealed particle filter (APF). In both methods Gaussian diffusion as dynamics model was applied. The performance evaluation is measured using real-life benchmark dataset *HumanEva* [11]. The motion sequence is synchronized with measurements from the *MOCAP* system and thus it is possible to evaluate the difference between the true values of pose configuration with the estimated ones using the following equation ($\mathbf{w}(\cdot) \in \mathcal{W}$ denotes M points on a body for given state variables): $\mathrm{err}(\hat{\mathbf{x}}_{1:T}) = \frac{1}{TM}\sum_{t=1}^{T}\sum_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}(\mathbf{x}_t) - \mathbf{w}(\hat{\mathbf{x}}_t)\|$. The obtained value of the error $\mathrm{err}(\hat{\mathbf{x}}_{1:T})$ is expressed in millimeters.

In the experiment we used two motion types, namely, *walking* and *jogging*, performed by three different persons, i.e., S1, S2, S3, which results in six various sequences. In each sequence we used 350 and 300 frames from different training trials for training and validation sets, respectively. Only the sequence S1-Jog contained 200 and 200 frames in training and validation sets, respectively. For testing we utilized first 200 frames from validation trial.

In the empirical study we used the following number of particles: (i) MRPF with 500 particles, (ii) PF with 500 particles, and (iii) APF with 5 annealing layers with 100 particles each. The low-dimensional manifold had 2 dimensions. All parameters (except $\gamma_{\mathbf{x}} = 10^{-4}$) were set according to the optimization process. The methods were initiated 5 times.

Results and Discussion. The averaged results obtained within the experiment are gathered in Table 1. The results show that the proposed approach with MRPF gave the best results except the sequence S1-Jog for which PF was slightly better. It is probably caused by the low-quality of this sequence which resulted in shorter training and validation sets. Because of this fact the manifold was not fully discovered.

The worst performance was obtained by the APF. The explanation for such result can be given as follows. First, the likelihood model used in the experiment is highly noised by the low quality of the silhouettes achieved in the background subtraction process. The noise in the likelihood model leads to displacement of extrema and thus wrong tracking. Second, the number of particles can be insufficient.

Table 1. The motion tracking errors err($\hat{\mathbf{x}}_{1:T}$) (in millimeters) for all methods are expressed as an average and a standard deviation (in brackets). The best results are in bold.

Sequence	APF	SIR	MRPF
S1-Walk	107(31)	82(18)	**69**(**7**)
S1-Jog	111(17)	**81**(**4**)	82(8)
S2-Walk	106(16)	95(7)	**86**(**12**)
S2-Jog	121(9)	106(13)	**94**(**8**)
S3-Walk	114(27)	88(13)	**79**(**10**)
S3-Jog	111(27)	117(29)	**70**(**8**)

In the summary, the manifold regularized particle filter seems to correctly follow the trajectory on the low-dimensional manifold. In other words, the prior knowledge about the low-dimensional manifold of the human pose configuration in motion is properly introduced. Additionally, MRPF obtained not only lower average error but also lower standard deviation in comparison to PF and APF. This result allows to presume that MRPF is more stable. However, in order to resolve this issue conclusively more experiments are needed.

7 Conclusions

In this paper, a fully Bayesian approach to the articulated human motion tracking was proposed. The modification of the particle filter technique is based on introducing low-dimensional manifold as a regularizer which incorporates prior knowledge about the specificity of human motion. The application of the low-dimensional manifold allows to restrict the space of possible pose configurations. The idea is based on the application of GPLVM with back constraints. At the end of the paper, the experiment was carried out using the real-life benchmark dataset *HumanEva*. The proposed approach was compared with two particle filters, namely, PF and APF, and the obtained results showed that it outperformed both of them.

References

1. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(1), 44–58 (2006)
2. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. International Journal of Computer Vision 61(2), 185–205 (2005)
3. Kanaujia, A., Sminchisescu, C., Metaxas, D.: Semi-supervised hierarchical models for 3D human pose reconstruction. In: CVPR 2007 Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007)
4. Lawrence, N.D.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. Journal of Machine Learning Research 6, 1783–1816 (2005)

5. Lawrence, N.D., Quiñonero-Candela, J.: Local distance preservation in the GP-LVM through back constraints. In: ICML 2006 Proceedings of the 23rd International Conference on Machine Learning, pp. 513–520 (2006)
6. Li, R., Tian, T., Sclaroff, S., Yang, M.: 3D human motion tracking with a coordinated mixture of factor analyzers. International Journal of Computer Vision 87, 170–190 (2010)
7. Memisevic, R., Sigal, L., Fleet, D.J.: Shared kernel information embedding for discriminative inference. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(4), 778–790 (2012)
8. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90–126 (2006)
9. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108, 4–18 (2007)
10. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
11. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision 87(1), 4–27 (2010)
12. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M.: Tracking loose-limbed people. In: CVPR 2004 Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (2004)
13. Taylor, G.W., Sigal, L., Fleet, D.J., Hinton, G.E.: Dynamical binary latent variable models for 3D human pose tracking. In: CVPR 2010 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (2010)
14. Tian, T., Li, R., Sclaroff, S.: Tracking human body pose on a learned smooth space. Technical Report 2005-029, Boston University Computer Science Department (2005)
15. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR 2011 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (2011)
16. Wang, J., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), 283–298 (2008)

Associative Learning Using Ising-Like Model

Jakub M. Tomczak

Institute of Computer Science, Wrocław University of Technology,
wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
jakub.tomczak@pwr.wroc.pl

Abstract. In this paper, a new computational model of associative learning is proposed, which is based on the Ising model. Application of the stochastic gradient descent algorithm to the proposed model yields an on-line learning rule. Next, it is shown that the obtained new learning rule generalizes two well-known learning rules, i.e., the Hebbian rule and the Oja's rule. Later, the fashion of incorporating the cognitive account into the obtained associative learning rule is proposed. At the end of the paper, experiments were carried out for testing the backward blocking and the reduced overshadowing and blocking phenomena. The obtained results are discussed and conclusions are drawn.

Keywords: associative learning, Ising model, energy-based model, Hebbian rule, Oja's rule, Rescorla-Wagner model, backward blocking, reduced overshadowing.

1 Introduction

Understanding animal or human learning process remains the most intriguing question in psychology, artificial intelligence, and cognitive science [13,19]. There are different approaches to understand learning processes. According to the Marr's tri-level hypothesis [11], the learning process can be considered at the representational level, i.e., by studying neurobiological properties of the brain [2], or at the algorithmic level, for example, by investigating how many information humans can process [6], or at the computational level which aims at describing what information are processed and what abstract representation has to be used in order to solve the learning problem [22]. In this work, we focus on modelling learning process at the computational level.

In the literature, two prominent types of accounts have been offered to explain the learning process phenomenon at the computational level [19,20], namely, associative (contingency) perspective, and cognitive (inferential, propositional) perspective. The first one tries to explain learning process as an unconscious process in which associations are built between cues (stimuli) and outcome (target) [5,18]. There are different models which are based on the associative account, e.g., classical Hebbian rule [7] and Rescorla-Wagner model [18], and more recent probabilistic models which use Kalman filter [5,9] or noisy logic gate [9]. In the second approach it is claimed that an explicit reasoning process leads to

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_28, © Springer International Publishing Switzerland 2014

inferences about causal relations between stimuli and target [15]. Recently, there were a series of probabilistic models in which the human inference is explained in terms of Bayesian learning paradigm [22].

In this paper, we concentrate on the associative account (an unaware learning phase) and try to combine it with the cognitive perspective (an aware learning phase). Such approach seems to be a reasonable direction because recent studies show that both types of learning processes co-exist [20]. Our model of associative learning is based on the Ising model [4,12]. The Ising model, originally developed in the statistical physics, is an energy-based model which associates an energy (a real value) with a system's state. This model has been successfully applied to many real-life problems, e.g., image de-noising [3], opinion evolution in closed community (also known as Sznajd model) [21], or biophysical dynamics modelling [10]. Moreover, the Ising model is a starting point in many models used in the field of machine learning, e.g., Hopfield networks [8] or Boltzmann machines [1].

The contribution of the paper is threefold:

1. A new computational model for associative learning basing on the Ising model is proposed. The model leads to an on-line learning rule which generalizes the well-known Hebbian rule and Oja's rule.
2. The proposed approach is qualitatively evaluated on benchmark experiments, namely, *backward blocking* and *reduced overshadowing and blocking*.
3. A fashion of incorporating cognitive perspective into the associative learning rule is outlined.

The paper is organized as follows. In Section 2 first the problem of associative learning is stated, and then the classical Rescorla-Wagner model is given (Section 2.1), and the proposed model is described (Section 2.2). Additionally, the generalizations of the well-known learning rules are derived. In Section 3 the experiments are carried out: backward blocking (Section 3.1), and reduced overshadowing and blocking (Section 3.2). Next, the results are discussed in Section 3.3. At the end, the conclusions are drawn in Section 4.

2 Associative Learning

The aim of a learning process is to find a dependency between cues (stimuli) and outcomes (responses or targets). The associative account for learning focuses on analysing links strengthening between stimuli and responses. In other words, it is the study of how animals or humans learn predictive relationships.

We distinguish a vector of D cues, $\mathbf{x} \in \{0,1\}^D$, and a single outcome, $t \in \{-1,1\}$.[1] If the d^{th} cue is present, we write $x_d = 1$, and $x_d = 0$ – otherwise. Additionally, we denote learner's knowledge about the associative strength between d^{th} cue and the outcome as w_d. Hence, we have a vector of D associative weights, $\mathbf{w} \in \mathbb{R}^D$.

[1] In this paper, we decided on coding the outcome using -1 and $+1$ because we want to differentiate the negative and positive responses, respectively.

The goal of the associative learning is to determine the weights values for given N observations, $\mathcal{D} = \{(\mathbf{x}_n, t_n)\}_{n=1}^{N}$. Positive values of weights represent strong relationship while negative values of weights – weak or none influence of cues on the outcome.

In further considerations we assume a model in which the outcome is a weighted sum of the cue activations:

$$y(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x}. \tag{1}$$

The meaning of the model is straightforward – each cue contributes to the outcome with its associative strength.

2.1 Classical Approach – Rescorla-Wagner Model

In the classical approach to the associative learning one aims at minimizing the error between the true target value and the predicted one, which is a function in the following form:

$$\mathcal{E}(\mathbf{w}) = \sum_{n=1}^{N} \mathcal{E}_n(\mathbf{w}), \tag{2}$$

where

$$\mathcal{E}_n(\mathbf{w}) = \big(t_n - y(\mathbf{x}_n; \mathbf{w})\big)^2.$$

In general, the stochastic (on-line) gradient descent (SGD) takes the following form ($\eta > 0$ – learning parameter, $\nabla_{\mathbf{w}}$ – gradient operator over $\mathbf{w}$):

$$\mathbf{w} := \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{E}_n(\mathbf{w}), \tag{3}$$

Applying SGD to the error function (2) yields the following learning rule:

$$\mathbf{w} := \mathbf{w} + \eta (t_n - \mathbf{w}^\top \mathbf{x}_n) \mathbf{x}_n, \tag{4}$$

which is known as *Rescorla-Wagner model* in the psychology domain [14,18].

2.2 Our Approach – Ising-Like Model

The Energy Function. In this paper, we propose a model which is driven by other motivation than simple error minimization approach. We want to associate an *energy* (a real value) with current learner's knowledge (state), i.e., weights' values, by proposing *an energy function*. We formulate the energy function as follows.

We know that there must be a strong correlation between the true outcome t and the model $y(\mathbf{x}; \mathbf{w})$. This dependency can be captured by the form $-c\mathbf{w}^\top \mathbf{x} t$, where $c > 0$. This has the desired effect of giving lower energy if $y(\mathbf{x}; \mathbf{w})$ and t have the same signs and higher energy – otherwise.

Additionally, we want to favour neighbouring cues to have the same signs, i.e., either both positive or both negative. We can do that by introducing a neighbourhood matrix $\mathbf{V} \in \{0,1\}^{D\times D}$ which determines the connections among cues. Thus, the energy associated with stimuli can be calculated using $-b\mathbf{w}^{\top}\mathbf{V}\mathbf{w}$, where $b \in \mathbb{R}$. Moreover, we can bias cues towards particular signs, i.e, positive (cues are present) or negative (cues are absent). This is equivalent to adding an extra term $-h\mathbf{w}$, where $h \in \mathbb{R}$.

Finally, we get the energy function in the following form:

$$E(\mathbf{w}) = -h\mathbf{w} - b\mathbf{w}^{\top}\mathbf{V}\mathbf{w} - c\mathbf{w}^{\top}\mathbf{x}t. \tag{5}$$

This is a modification of the *Ising model* [4,12], which has been widely studied in the statistical physics domain. The parameters h and b have certain interpretations in physics,[2] namely, h corresponds to the presence of an external magnetic field, b is the coupling between spins. The last parameter c was not introduced in the original Ising model for magnetic systems and thus has no clear interpretation. However, in our application it determines the correlation between the model and the outcome.

Learning Rule. Similarly to the Rescorla-Wagner model, in order to obtain the learning rule we want to apply SGD but to the energy function (5). Therefore, we need to calculate the gradient:[3]

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -h - b(\mathbf{V} + \mathbf{V}^{\top})\mathbf{w} - c\mathbf{x}t. \tag{6}$$

Finally, we get the following learning rule:

$$\mathbf{w} := \mathbf{w} + \eta(h + b(\mathbf{V} + \mathbf{V}^{\top})\mathbf{w} + c\mathbf{x}t) \tag{7}$$

Let us notice that the learning rule (7) is a generalization of two well-known learning rules. First, for $h = 0$ $b = 0$, and $c = 1$ we get:

$$\mathbf{w} := \mathbf{w} + \eta\mathbf{x}t, \tag{8}$$

which is known as *Hebbian rule* [7].

Second, for $h = 0$, $b = -0.5$, $c = 1$, and for the neighbourhood matrix $\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix, we get:[4]

$$\mathbf{w} := \mathbf{w} + \eta(\mathbf{x}t - \mathbf{w}), \tag{9}$$

which is known as *Oja's rule* [16].

[2] Originally, the Ising model has been proposed to investigate properties of idealized magnetic systems. In the Ising model there is a lattice of spins which can take only two values -1 and $+1$. Here, we talk about associative weights which are real-valued.

[3] We use the following property (equation (81) on page 11 in [17]):

$$\frac{\partial \mathbf{x}^{\top}\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^{\top})\mathbf{x}.$$

[4] In fact, in order to obtain the original Oja's rule we should write

$$\mathbf{w} := \mathbf{w} + \eta t(\mathbf{x} - \mathbf{w}t)$$

but since $t \in \{-1, 1\}$, and thus $t^2 = 1$, we omit writing t^2 in the final equation.

Combining Associative and Cognitive Approaches. In recent studies it has been suggested that associative and cognitive accounts should be combined rather than considered apart [20]. In this paper, we focus on associative learning but here we point out a possible direction of how to incorporate cognitive perspective into the associative one.

We introduce a new concept which we call *surprise factor*, and define it as an absolute difference between the true target and the model's outcome:

$$s = |t - \mathbf{w}^{\top}\mathbf{x}|. \tag{10}$$

The surprise factor represents the inferential account of a new observation, i.e., the bigger predictive mistake learner makes, the bigger influence should have the observation on the learner. In other words, if the target and the model's outcome are different, the surprise factor reflects the surprise level of the learner about the considered observation.

The proposed quantity can be used to modify the learning rule by changing the learning rate, i.e., $\eta(s) = \eta/s$. In case of observations which are correctly predicted, the learner is not surprised by data and thus makes small contribution to her knowledge. Here we see that the learner has to perform simple cognitive process represented by calculating the surprise factor and then apply the typical associative learning rule with the modified learning rate.

The surprise factor is a simple representation of a cognitive process but we can think of more sophisticated ones. For example, we can try to propose a model of searching in the memory for similar instances and then modify the learning rate. In general, we can formulate an inferential model of the cognitive process which reflects influence of awareness on animal or human perception of the world.

3 Experiments

In the psychology domain there were series of designed experiments which served as benchmark test-cases in evaluating and understanding animal or human learning process [19]. There are two typical experiments which are used in a qualitative comparison of computational models of learning, namely, *backward blocking* [9] and *reduced overshadowing and blocking* [9,15].

In the following sections the proposed approaches (with parameters: $h = -0.05$, $b = -0.2$, $c = 1$, $\eta = 0.1$, and the neighbourhood matrix $\mathbf{V}$ – an upper triangular matrix of ones), that is,

- without cognitive account (in the experiment called *Our model*);
- with the cognitive account ((in the experiment called *Our model + cognition*),

are compared with the classical learning models (with learning rate equal $\eta = 0.1$):

- the Rescorla-Wagner model;
- Hebbian rule;
- Oja's rule.

Table 1. Structure of the learning process for the backward blocking experiment

Stage	Frequency	x_1	x_2	t
I	10	1	1	1
II	10	1	0	1

The parameters are set arbitrarily. We will not discuss the parameters' influence on the model's performance. We leave this issue for further research.

3.1 Backward Blocking

In the backward blocking experiment there are two stimuli, $\mathbf{x} \in \{0, 1\}^2$, and one outcome, $t \in \{-1, 1\}$. The learning process consists of two training stages (see Table 1). The first phase of training comprises of 10 observations in which both cues occur with the positive outcome. The second stage of training has 10 trials in which only x_1 occurs along with the positive response. The observed phenomenon is that when x_2 is tested by itself at the end of the second stage it evokes lower associative strength than at the end of the first phase.

The results obtained for the proposed approaches and the three classical learning models are presented in the Figure 1. The associative weights' values are given for learning stage I and II (the learning stage number is denoted in the brackets).

3.2 Reduced Overshadowing and Blocking

In the reduced overshadowing with blocking experiment there are four stimuli, $\mathbf{x} \in \{0, 1\}^4$, and one outcome, $t \in \{-1, 1\}$. The learning process consists of two training stages (see Table 2). The first phase of training comprises of 6 observations in which only x_1 occurs with the positive outcome and then 6 observations in which x_3 occurs with the negative outcome. The second stage of training has 6 observations in which x_1 and x_2 occur along with the positive response and next 6 observations in which x_3 and x_4 occur with the positive response. The observed phenomenon is that first x_3 entails no outcome and thus it reduces overshadowing when x_3 occurs together with x_4. Additionally, at the end of the learning process the associative strength of x_2 should be less than x_4 because the x_2 is blocked by x_1.

Table 2. Structure of the learning process for the reduced overshadowing and blocking experiment

Stage	Frequency	x_1	x_2	x_3	x_4	t
Ia	6	1	0	0	0	1
Ib	6	0	0	1	0	-1
IIa	6	1	1	0	0	1
IIb	6	0	0	1	1	1

The results obtained for the proposed approaches and the three classical learning models are presented in the Figure 2. The associative weights' values are given for learning stage Ia, Ib and IIa, IIb (the learning stage number is denoted in the brackets).

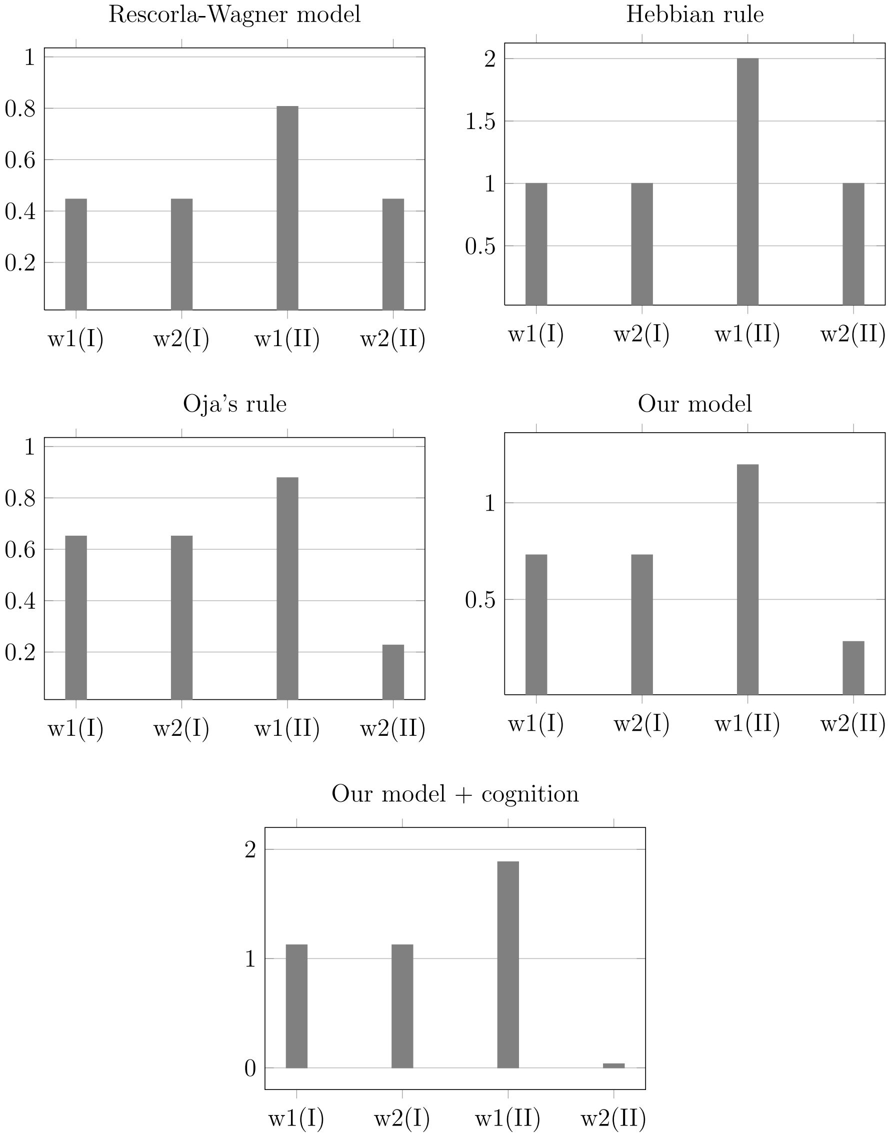


Fig. 1. Summary performance comparison of all considered in the paper learning rules in the backward blocking experiment

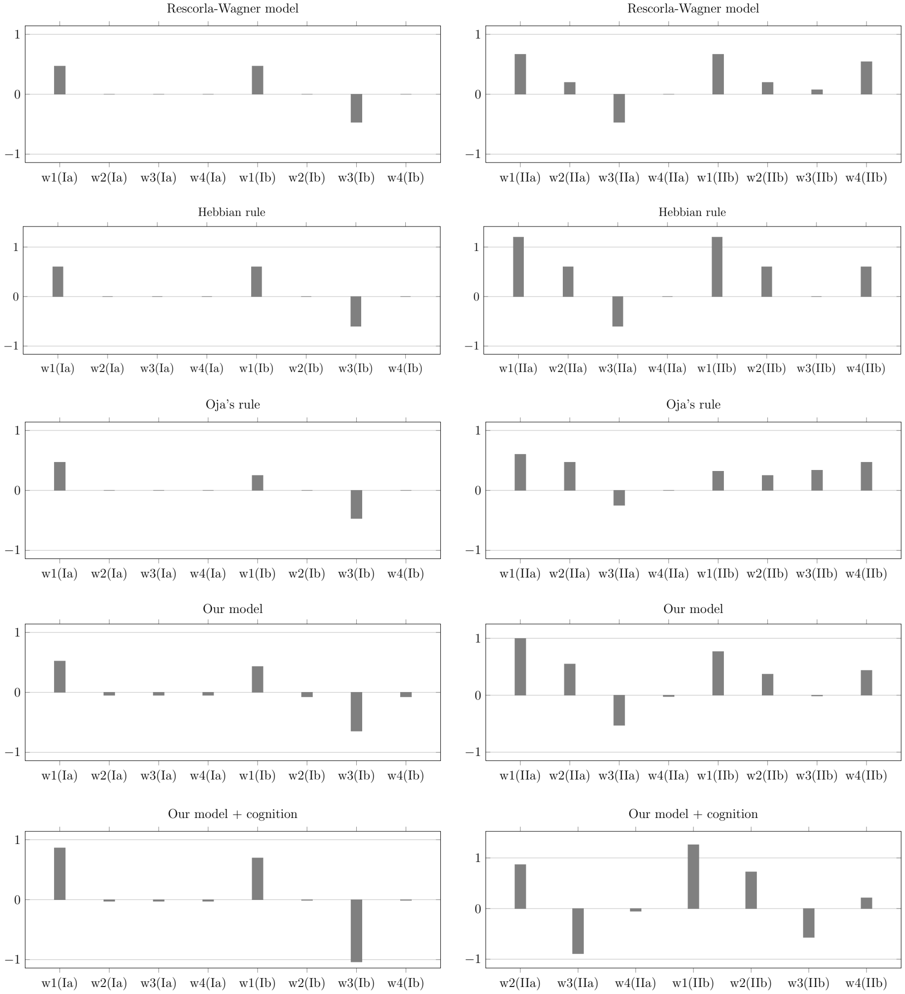


Fig. 2. Summary performance comparison of all considered in the paper learning rules in the reduced overshadowing and blocking experiment

3.3 Discussion

In the first experiment, the obtained results (see Figure 1) show that the Rescorla-Wagner model does not exhibit backward blocking which is a known fact [14]. Similarly, the Hebbian rule simply leads to accumulating the number of co-occurrences of cues and thus it is insensitive to the backward blocking phenomenon. On the other hand, the Oja's rule and both of our proposed models are able to model the backward blocking. This result indicates the importance of introducing $\mathbf{w}^\top \mathbf{V} \mathbf{w}$. In fact, this part of the energy function plays a role of a *regularizer* if we notice that $\mathbf{w}^\top \mathbf{V}\mathbf{w} = \|\mathbf{w}\|^2_{\mathbf{V}}$. [5] In other words, it tries to *pull* all weights' values towards zeros and therefore the Oja's rule and our approaches exhibit backward blocking.

In the second experiment, the obtained results (see Figure 2) are quite surprising. The Rescorla-Wagner model exhibit both reduced overshadowing and blocking. However, our proposition without and with cognition shows more evidently these both phenomena. It is important that our model indicates strong associative weights for cues 1, 2, and 4, and zero or strong negative (for our proposition with cognition) association for cue 3. At the end, let us notice that Hebbian rule failed completely. The Oja's rule performed better but it put too much weight on last observations and thus the cue 3 and cue 4 have too strong associative strengths.

4 Conclusions

In this paper, the new computational model of associative learning was proposed. It is based on the Ising model which has been successfully applied in the statistical physics [12] and other domains [3,10,21]. It was shown that the obtained new learning rule (7) generalizes the Hebbian rule (8) and the Oja's rule (9). Next, the fashion of incorporating the cognitive account into the obtained associative learning rule is proposed. Possibly, this indicates a new direction for future research. At the end of the paper, experiments were carried out for testing the backward blocking and the reduced overshadowing and blocking phenomena. The obtained results have revealed the supremacy of our model over the classical approaches to the associative learning.

In the ongoing research we develop the probabilistic approach to the presented problem, i.e., application of Boltzmann machine [1]. In this paper, we have pointed out the possibility of combining associative and cognitive accounts. Considering the experimental results this idea seems to be an interesting direction for further investigations. Additionally, we have assumed an arbitrary values of parameters of the model. The parameters' values should be fitted to individuals basing on real data. However, we leave investigating this aspect as future research.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines. Cognitive Science 9(1), 147–169 (1985)
2. Ashby, F.G., O'Brien, J.B.: Category learning and multiple memory systems. Trends in Cognitive Sciences 9(2), 83–89 (2005)
3. Bishop, C.M.: Pattern recognition and machine learning. Springer, New York (2006)

[5] Since $\mathbf{V}$ is an upper triangular matrix of ones, its all eigenvalues are positive and hence it is a positive definite matrix.

4. Cipra, B.A.: An introduction to the Ising model. American Mathematical Monthly 94(10), 937–959 (1987)
5. Dayan, P., Kakade, S., Montague, P.R.: Learning and selective attention. Nature Neuroscience 3, 1218–1223 (2000)
6. Halford, G.S., Baker, R., McCredden, J.E., Bain, J.D.: How many variables can humans process? Psychological Science 16(1), 70–76 (2005)
7. Hebb, D.O.: The Organization of Behavior. Wiley & Sons, New York (1949)
8. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences 79(8), 2554–2558 (1982)
9. Kruschke, J.K.: Bayesian approaches to associative learning: From passive to active learning. Learning & Behavior 36(3), 210–226 (2008)
10. Lis, M., Pintal, L., Swiatek, J., Cwiklik, L.: GPU-Based Massive Parallel Kawasaki Kinetics in the Dynamic Monte Carlo Simulations of Lipid Nanodomains. Journal of Chemical Theory and Computation 8(11), 4758–4765 (2012)
11. Marr, D.: Vision. W.H. Freeman, San Fransisco (1982)
12. MacKay, D.J.: Information theory. inference and learning algorithms. Cambridge University Press (2003)
13. McClelland, J.L.: Is a machine realization of truly human-like intelligence achievable? Cognitive Computation 1(1), 17–21 (2009)
14. Miller, R.R., Barnet, R.C., Grahame, N.J.: Assessment of the Rescorla-Wagner model. Psychological Bulletin 117(3), 363–386 (1995)
15. Mitchell, C.J., De Houwer, J., Lovibond, P.F.: The propositional nature of human associative learning. Behavioral and Brain Sciences 32(02), 183–198 (2009)
16. Oja, E.: Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology 15(3), 267–273 (1982)
17. Petersen, K.B., Pedersen, M.S.: The matrix cookbook (2012)
18. Rescorla, R.A., Wagner, A.R.: A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II: Current Research and Theory, pp. 64–99 (1972)
19. Shanks, D.R.: Learning: From association to cognition. Annual Review of Psychology 61, 273–301 (2010)
20. Sternberg, D.A., McClelland, J.L.: Two mechanisms of human contingency learning. Psychological Science 23(1), 59–68 (2012)
21. Sznajd-Weron, K., Sznajd, J.: Opinion evolution in closed community. International Journal of Modern Physics C 11(06), 1157–1165 (2000)
22. Tenenbaum, J.B., Griffiths, T.L., Kemp, C.: Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences 10(7), 309–318 (2006)

Cost Sensitive SVM with Non-informative Examples Elimination for Imbalanced Postoperative Risk Management Problem

Maciej Zięba, Jerzy Świątek, and Marek Lubicz

Faculty of Computer Science and Management, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

Abstract. In this paper we propose a novel combined approach to solve the imbalanced data issue in the application to the problem of the post-operative life expectancy prediction for the lung cancer patients. This solution makes use of undersampling techniques together with cost-sensitive SVM (Support Vector Machines). First, we eliminate non-informative examples by applying Tomek links together with one-sided selection. Second, we take advantage of using cost-sensitive SVM with penalty costs calculated respecting cardinalities of minority and majority examples. We evaluate the presented solution by comparing the performance of our method with SVM-based approaches that deal with uneven data. The experimental evaluation was performed on real-life data from the postoperative risk management domain.

1 Introduction

The problem of imbalanced data is widely observed in the classification domain when cardinalities of examples in the training set are significantly different. This phenomenon is mainly observed in the binary classification problems where we can distinguish minority and majority class [10]. The minority examples (assumed to be positives) are the rarely observed instances which are often much more significant in the context of a decision problem than the majority observations (negatives) that dominate the training data.

The imbalanced data has an impact on the training process in which the constructed classifier is biased toward majority class. In practice, despite high accuracy achieved by the classifier, this decision model is ineffective in detecting minority examples. In the extreme cases the trained classifier imitates *Zero Rule* method which assigns only majority class, independently on the vector of features.

Typical training methods used to train the classifiers fail in constructing decision models from imbalanced data. For instance, applying methods for constructing decision trees is not recommended because of the pruning mechanism which eliminates the weakly supported components of the tree associates with the minority class. Statistical approaches, like *Naive Bayes* method, have tendency to classify examples to the majority class because of the biased *a priori*

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_29, © Springer International Publishing Switzerland 2014

estimator. Using a typical *SVM* (*Support Vector Machines*) model on hardly separable, imbalanced data may lead in shifting the margin of separation toward the minority class. Therefore, there is a need of constructing training algorithms which deal with the problem of uneven data.

The problem of imbalanced data is present in many applications, including: biomedical systems, financial decision making systems, supervised outlier detection and many others. In this paper we concentrate on the problem of postoperative life expectancy prediction, which is one of the key issues in the surgery risk management field. We define the problem in terms of classification, where the minority class represents the situation in which a patient died in one year after the operation and the majority class is combined with the fact that the person survived the given period. Due to the fact that deaths are observed rarely comparing to survivals, the problem of constructing the classifier is the typical problem of learning from the imbalanced data and seeks the solutions out of the typical approaches.

To solve the stated manner for the described classification problem we propose to use cost-sensitive *SVM* enriched by preprocessing the methods for non-informative examples elimination. In our approach we suggest to combine the benefits of using so-called *Tomek links* together with the modified algorithm of one sided selection to eliminate the redundant instances. Next, we present how to modify the criterion of learning *SVM* by incorporating different penalty cost values for positive and negative examples to deal with disproportions in the training data. Finally, we compare the results gained by our approach with the other solutions dedicated for the problem of learning from the imbalanced data using real-life dataset from the risk management domain.

The paper is organized as follows. In Section 2 we present typical approaches for the problem of uneven data described in literature. In Section 3 we describe how to eliminate noise examples from the training data before constructing the learner. Section 4 presents cost-sensitive approach for learning *SVM* to solve the problem of imbalance data. The experimental results are included in Section 5. Section 6 summarizes this work with some conclusions.

2 Related Works

The problem of imbalanced data is well-studied and many solutions have been proposed in the literature. In general, there are three groups of approaches to solve the problem of uneven data [8,10]:

- external methods,
- internal methods,
- cost-sensitive approaches.

The methods aggregated in the first group solve the problem of uneven data on the preprocessing stage, independently on the training procedure. The problem of disproportions in the training set is eliminated by generating additional artificial examples from the minority class (oversampling), or by eliminating redundant observations from the majority class (undersampling). One of the most commonly used oversampling solutions is Synthetic Minority Over-sampling TEchnique (SMOTE) presented in [4]. The main idea of this method is to generate artificial samples in the close neighbourhood of the minority examples to increase the impact of positives in the training process. SMOTE is often used as a component of more sophisticated, internal approaches. On the other hand, some redundant examples can be eliminated to make training data more balanced. It is usually performed by applying random undersampling, or by making use of $K - NN$ based approaches, like *Tomek links* [1,17], or one-sided selection technique [11].

The internal approaches solve the problem of imbalanced data directly during the training procedure. In this group we distinguish methods that combine external approaches with the ensemble classification models. The main idea of such solutions is to construct the base classifiers on the training set modified using some of the external approaches. This kind of methods make use either of *bagging* (*QuasiBagging* [2], *SMOTEBagging* [19]), or *boosting* (*SMOTEBoost* [3], *RAMOBoost* [5], *DataBoost-IM* [9]). The other group of the internal approaches makes use of *granular computing* techniques [16], the main idea of which is to make knowledge-oriented decomposition of the main problem into parallel, balanced sub-problems, named *information granules*. The other important group of methods takes advantage of active learning solutions [7], the main idea of which is to select the informative examples in the neighbourhood of the borderline between two classes.

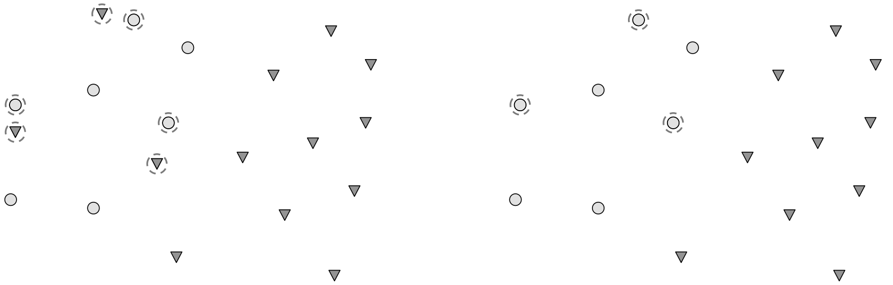
The last group of methods assigns weight values to the examples from the training data in such way, that the minority examples become more significant in training data than the objects from the majority class. Typical training methods are enriched by the mechanisms that include the information about the importance of each example in data. Such modifications are successively applied in *boosting*-based approaches [15], decision trees [6], neural networks [12] and *SVM*s [13,18].

3 Eliminating Redundant Examples

In this work we propose the complex approach to deal with imbalanced data in which we can distinguish two stages:

- *Preprocessing stage.* At this stage we eliminate the examples from the majority class by applying two techniques for clearing non-informative examples.
- *Training stage.* At this stage we propose to use cost-sensitive SVM with the modified learning criterion.

In this work we recommend to use cost-sensitive approach in which the cost values are calculated basing on the cardinalities of examples from majority and minority classes. Such approach can be applied if the data set is composed only of the informative examples. To solve the stated manner we propose to use two techniques of eliminating redundant members of training set. First, we propose to apply *Tomek links* technique which aims at detecting "noise examples" from the majority class. Secondly, we apply the one-sided selection method to keep only examples from the majority class that are situated in the neighbourhood of the minority class instances.



(a) Examples associated using *Tomek* links (in the circles).

(b) Examples after elimination of majority examples in *Tomek* relation.

Fig. 1. Using *Tomek links* for eliminating redundant examples from majority class

Tomek link [17] is a kind of relation of two examples, $\mathbf{x}_n$ and $\mathbf{x}_m$, from different classes ($y_n \neq y_m$), where $\mathbf{x}_n$ is the nearest neighbour of $\mathbf{x}_m$ and $\mathbf{x}_m$ is the closest neighbour of $\mathbf{x}_n$. Formally, there is no such example $\mathbf{x}_l$, for which $d(\mathbf{x}_m, \mathbf{x}_l) < d(\mathbf{x}_m, \mathbf{x}_n)$, or $d(\mathbf{x}_n, \mathbf{x}_l) < d(\mathbf{x}_n, \mathbf{x}_m)$. Distance measure $d(\cdot,\cdot)$ is compatible with feature space considered in classification process and can be described with the following equation:

$$\begin{aligned} d(\mathbf{x}_n, \mathbf{x}_m) = ||\phi(\mathbf{x}_n) - \phi(\mathbf{x}_m)||_2 &= ||\phi(\mathbf{x}_n)||_2 + ||\phi(\mathbf{x}_m)||_2 - 2\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) \\ &= K(\mathbf{x}_n, \mathbf{x}_n) + K(\mathbf{x}_m, \mathbf{x}_m) - 2K(\mathbf{x}_n, \mathbf{x}_m), \quad (1) \end{aligned}$$

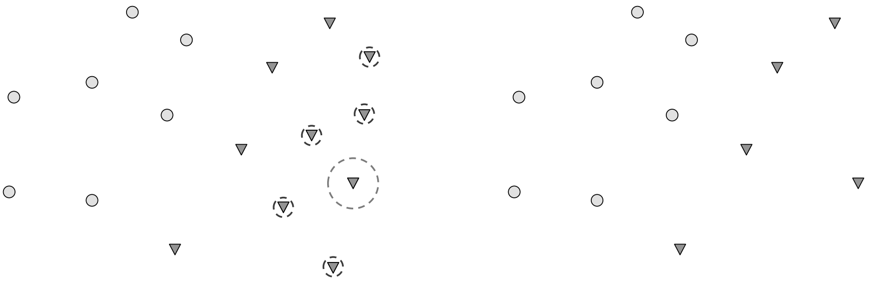
where $\phi(\cdot)$ is transformation to space where SVM makes linear separation and $K(\cdot,\cdot)$ is the corresponding kernel function used by the classifier. The process of eliminating examples from the majority class is simply performed by identifying the observations which are linked by *Tomek* relation (Figure 1 a)) and removing them from the training data (Figure 1 b)).

The second step of the preprocessing procedure is related to eliminating examples using one-sided approach [11]. The procedure of removing non-informative observations is presented by Algorithm 1. On the input we consider the dataset $\mathbb{S}_N$ composed of the pairs $(\mathbf{x}_n, y_n)$, where $\mathbf{x}_n$ represents the vector of features and y_n the corresponding class label ($y_n = -1$ if example is from minority class, $y_n = +1$ otherwise). In the first step we generate two data sets $\mathbb{S}_{N_+}$ and $\mathbb{S}_{N_-}$, where $\mathbb{S}_{N_+}$ is composed only of the positive examples and $\mathbb{S}_{N_-}$ contains only the negatives (N_+ denotes total number of examples from minority class and N_- from the majority class). Further, we randomly select a positive example from

Algorithm 1. One-sided selection algorithm.

```
Input  : Training set $\mathbb{S}_N = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
Output: Training set after elimination $\mathbb{S}_{N_2}$
$\mathbb{S}_{N_2} \longleftarrow \emptyset$;
$\mathbb{S}_{N_+} = \{(\mathbf{x}_n, y_n) \in \mathbb{S}_N : y_n = +1\}$;
$\mathbb{S}_{N_-} = \{(\mathbf{x}_n, y_n) \in \mathbb{S}_N : y_n = -1\}$;
Select $(\mathbf{x}_n, -1)$ randomly from $\mathbb{S}_{N_-}$;
$\mathbb{S}_{N_+} \longleftarrow \mathbb{S}_{N_+} \cup \{(\mathbf{x}_n, -1)\}$;
for $(\mathbf{x}_m, y_m, w_m^{(k)}) \in \mathbb{S}_{N_-}$ do
    $(\mathbf{x}_l, y_l) \longleftarrow \underset{(\mathbf{x}_l, y_l) \in \mathbb{S}_{N_+}}{\text{argmin}}\ d(\mathbf{x}_m, \mathbf{x}_l)$ ;
    if $y_l \neq y_m$ then
        $\mathbb{S}_{N_2} \longleftarrow \mathbb{S}_{N_2} \cup \{(\mathbf{x}_l, y_l)\}$;
    end
end
$\mathbb{S}_{N_2} \longleftarrow \mathbb{S}_{N_2} \cup \mathbb{S}_{N_+}$;
```

$\mathbb{S}_{N_-}$ and insert it into $\mathbb{S}_{N_+}$. Finally, we examine each of the examples taken from $\mathbb{S}_{N_-}$ and identify the closest neighbour of the instance from $\mathbb{S}_{N_+}$. If the closest neighbour belongs to the minority class, the selected instance is inserted into set $\mathbb{S}_{N_+}$. After examining all elements from $\mathbb{S}_{N_-}$ we return $\mathbb{S}_{N_+}$ as the reduced output dataset.



(a) Randomly selected example from majority class (big circle) and instances identified for elimination (small circles).

(b) Examples after applying undersampling procedure.

Fig. 2. The process of undersampling examples with one-sided method

The policy of one-sided elimination is based on the assumption that the non-informative examples are situated so far from the borderline between two classes, that even the randomly selected example from the majority class is situated closer than any of the positives (Figure 2).

4 Cost-Sensitive SVM for Imbalanced Data

The presented methods of dealing with the problem of imbalanced data on the preprocessing stage reduce the number of redundant majority examples, but do not solve the issue on the satisfactory level. Therefore, we propose to use cost-sensitive SVM to solve the stated manner.

The problem of learning traditional SVM can be formulated as the optimization task in which we minimize the following criterion: [1]

$$Q(\mathbf{a}) = \frac{1}{2}\mathbf{a}^T\mathbf{a} + C\sum_{n=1}^{N}\xi_n, \tag{2}$$

under following conditions:

$$y_n(\mathbf{a}^T\phi(\mathbf{x}) + b) \geq 1 - \xi_n, \tag{3}$$

for $n \in \{1, \ldots, N\}$, where $\mathbf{a}$, b are the parameters of the linear classifier in the feature space determined by $\phi(\cdot)$, ξ_n is the *slack variable* that weakens the optimization criterion by introducing the *soft margin* and C is penalty parameter for non-zero ξ_n values. In this work we propose to use a modified criterion discussed in [13]:

$$Q(a) = \frac{1}{2}\mathbf{a}^T\mathbf{a} + C_+\sum_{n_+\in\mathbb{N}_+}\xi_{n_+}^k + C_-\sum_{n_-\in\mathbb{N}_-}\xi_{n_-}^k, \tag{4}$$

where $\mathbb{N}_+ = \{n \in \{1, \ldots, N\} : y_n = +1\}$, $\mathbb{N}_- = \{n \in \{1, \ldots, N\} : y_n = -1\}$ and C_+, C_- are the penalty parameters for positive and negative examples, respectively. We propose to calculate C_+ and C_- using the following formulas:

$$C_+ = C\frac{N}{2N_+}, \tag{5}$$

and

$$C_- = C\frac{N}{2N_-}. \tag{6}$$

We make use of cardinalities of majority and minority examples to increase the strength of penalization for incorrectly classified minority cases at the expense of majority instances. Applying data clearing technique eliminates the non-informative examples so the proportion $\frac{N_+}{N_-}$ is similar to the class ratio observed near the borderline. Additionally, the generalization of the SVM is controlled by C parameter.

The dual form of the learning criterion is described by the equation:

$$Q_D(\boldsymbol{\lambda}) = \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N}\lambda_i\lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{7}$$

[1] The nominal attributes are transformed to binary features.

where $\boldsymbol{\lambda}$ is the N-element vector of Lagrange's multipliers. The problem of learning SVM in dual form is formulated as the optimization task in which we are interested in finding the maximal value of (7) under the following conditions:

$$0 \leq \lambda_n \leq C_+ \quad n \in \mathbb{N}_+, \tag{8}$$

$$0 \leq \lambda_m \leq C_- \quad m \in \mathbb{N}_-. \tag{9}$$

The stated problem is identified as a convex optimization task. Due to large number of the multipliers, various solutions are used to solve the given manner. The most common approach is Platt's *SMO* (Sequential Minimal Optimization) [14], which optimizes only two multipliers in each iteration of the heuristic. Classical *SMO* is used to solve the problem of learning balanced *SVM*, but it can be easily adjusted to the imbalanced issue, by proper selection of C_+ and C_- instead of C.

5 Experimental Results

In this work we consider the application of the presented approach to the problem of the post-operative life expectancy prediction in the lung cancer patients. We evaluate the presented approach on real-life clinical data from Thoracic Surgery domain. The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for over 1200 consecutive patients who underwent major lung resections for primary lung cancer in the years 2007-2011. Each of the patients is described by the vector composed of 32 features representing his condition before, during and after the surgery. The problem was formulated as a binary classification task, in which the positive class represented deaths in one year after surgery (220 examples), and negative class, if the patient survived after surgery (983 cases). In such problem formulation we identified the imbalanced data phenomenon biased toward the negative class.

In the experiment we consider following methods to be tested for the selected decision problem:

- *SVM* trained with *SMO* (**SVM**).
- *SVM* with *SMOTE* sampling technique on preprocessing stage (**SMOTE + SVM**).
- *SVM* with non-informative examples elimination using the procedure described in Section 3 (**NE + SVM**).
- Cost-sensitive *SVM* for the imbalanced data presented in Section 4 (**C-SVM**).
- Cost-sensitive *SVM* with *SMOTE* sampling technique on the preprocessing stage (**SMOTE + CSVM**).
- Cost-sensitive *SVM* with the non-informative examples elimination (**NE + CSVM**).

Table 1. The confusion matrix for two-class decision problem

	Classified to positive class	**Classified to negative class**
Is member of positive class	TP (*True positive*)	FN (*False negative*)
Is member of negative class	FP (*False positive*)	TN (*True negative*)

We applied 5 folds cross validation as the testing methodology. The results of the experiment are given in Table 2. As the main criterion of evaluation we used *Gmean value*, which is typically used to evaluate the methods for imbalanced data:

$$GMean = \sqrt{TP_{rate} \cdot TN_{rate}}, \tag{10}$$

where TN_{rate} is named *specificity* rate, or simply TN (*ang. TN rate*) and is defined in the following way:

$$TN_{rate} = \frac{TN}{TN + FP}, \tag{11}$$

and TP_{rate} is called *sensitivity* rate, or TP (*ang. TP rate*), and is given by the equation:

$$TP_{rate} = \frac{TP}{TP + FN} \tag{12}$$

Values of TP (*ang. true positive*), FN (*ang. false negative*), FP (*ang. false positive*), TN (*ang. true negative*), are components of the confusion matrix given in Table 1).

Table 2. Results for thoracic surgery data

Method	$\mathbf{TP_{rate}}$	$\mathbf{TN_{rate}}$	**Accuracy**	**Gmean**
SVM	0.0818	**0.9827**	**0.8180**	0.2836
SMOTE + SVM	0.2182	0.8290	0.7174	0.4253
NE + SVM	0.2223	0.8281	0.7174	0.4295
C-SVM	0.5773	0.7101	0.6858	0.6402
SMOTE + C-SVM	0.2545	0.8179	0.7149	0.4563
NE + C-SVM	**0.6045**	0.7111	0.6916	**0.6556**

The lowest *GMean* value was achieved by *SVM* without any mechanisms of dealing with the imbalanced data problem. The classifier was strongly biased toward the majority class and successively detected only 8% of positive examples. Applying the preprocessing techniques together with *SVM* increased the *GMean* value, but the accuracy of detecting minority examples was still on the unsatisfactory level (21% – 23% of detected positives). Worthwhile results were observed, when cost-sensitive SVM was applied to the problem. Increasing the number of positives by generating artificial examples using $SMOTE$ decreased the *GMean* value significantly. It was mainly caused by the nature of SMOTE sampling technique which generates large set of non-informative examples. In such situation the cardinalities of examples from majority and minority class are comparable, while the examples cumulated near the borderline are still biased toward majority class. On the other hand, applying the methods for detecting non-informative examples resulted in the increase of TP_{rate} without any loss on TN_{rate}. For this approach we observed the highest value of *GMean* criterion. The proposed approach increased the *GMean* value mainly due to better estimation of C_+ and C_- in which only informative examples from training data were taken under consideration.

6 Conclusion

In this paper we propose the combined approach that make use of inner and outer mechanisms of dealing with the imbalanced data in application to the problem of the post-operative life expectancy prediction in the lung cancer patients. In this solution we eliminate the redundant instances from the majority class to increase the performance of cost-sensitive SVM. We presented experimental results on the real-life dataset which confirms the high quality of the combined solution.

Acknowledgments. The research by Maciej Zieba was co-financed by the European Union as part of the European Social Fund.

The research by Marek Lubicz was partly financed by the National Science Centre under the grant N N115 090939 "Models and Decisions in the Health Systems. Application of Operational Research and Information Technologies for Supporting Managerial Decisions in the Health Systems".

References

1. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6(1), 20–29 (2004)
2. Chang, E.Y., Li, B., Wu, G., Goh, K.: Statistical learning for effective visual information retrieval. In: Proceedings of the 2003 International Conference on Image Processing, vol. 3, pp. 609–613. IEEE (2003)
3. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O.: SMOTE: Synthetic Minority Oversampling TEchnique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
5. Chen, S., He, H., Garcia, E.A.: Ramoboost: Ranked minority oversampling in boosting. IEEE Transactions on Neural Networks 21(10), 1624–1642 (2010)

6. Elkan, C.: The foundations of cost-sensitive learning. In: The Proceedings of International Joint Conference on Artificial Intelligence, vol. 17, pp. 973–978. Lawrence Erlbaum Associates, Ltd. (2001)
7. Ertekin, S., Huang, J., Giles, C.L.: Active learning for class imbalance problem. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 823–824. ACM (2007)
8. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews 42(4), 3358–3378 (2012)
9. Guo, H., Viktor, H.L.: Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM SIGKDD Explorations Newsletter 6(1), 30–39 (2004)
10. He, H., Garcia, E.A.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)
11. Kubat, M., Matwin, S.: et al. Addressing the curse of imbalanced training sets: one-sided selection. In: ICML, pp. 179–186. Morgan Kaufmann Publishers (1997)
12. Kukar, M., Kononenko, I.: Cost-sensitive learning with neural networks. In: Proceedings of the 13th European Conference on Artificial Intelligence (ECAI 1998), pp. 445–449. Citeseer (1998)
13. Morik, K., Brockhausen, P., Joachims, T.: Combining statistical learning with a knowledge-based approach-a case study in intensive care monitoring. In: Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), pp. 268–277. Morgan Kaufmann (1999)
14. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge (1999)
15. Sun, Y., Kamel, M., Wong, A., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40(12), 3358–3378 (2007)
16. Tang, Y., Jin, B., Zhang, Y.Q.: Granular support vector machines with association rules mining for protein homology prediction. Artificial Intelligence in Medicine 35(1-2), 121–134 (2005)
17. Tomek, I.: Two Modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics 6(11), 769–772 (1976)
18. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the sensitivity of support vector machines. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCA I999), Workshop ML3, vol. 1999, pp. 55–60 (1999)
19. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining Proceedings, pp. 324–331. IEEE (2009)

Cost-Sensitive Extensions for Global Model Trees: Application in Loan Charge-Off Forecasting

Marcin Czajkowski[1], Monika Czerwonka[2], and Marek Kretowski[1]

[1] Faculty of Computer Science, Bialystok University of Technology,
Wiejska 45a, 15-351 Bialystok, Poland
{m.czajkowski,m.kretowski}@pb.edu.pl

[2] Collegium of Management and Finance, Warsaw School of Economics,
Al. Niepodleglosci 162, 02-554 Warsaw, Poland
monika.czerwonka@sgh.waw.pl

Abstract. Most of regression learning methods aim to reduce various metrics of prediction errors. However, in many real-life applications it is prediction cost, which should be minimized as the under-prediction and over-prediction errors have different consequences. In this paper, we show how to extend the evolutionary algorithm (EA) for global induction of model trees to achieve a cost-sensitive learner. We propose a new fitness function which allows minimization of the average misprediction cost and two specialized memetic operators that search for cost-sensitive regression models in the tree leaves. Experimental validation was performed with bank loan charge-off forecasting data which has asymmetric costs. Results show that Global Model Trees with the proposed extensions are able to effectively induce cost-sensitive model trees with average misprediction cost significantly lower than in popular post-hoc tuning methods.

Keywords: cost-sensitive regression, asymmetric costs, evolutionary algorithms, model trees, loan charge-off forecasting.

1 Introduction

In the vast number of contemporary systems, information including business, research and medical issues is collected and processed. In real-life data mining problems, the traditional minimization of prediction errors may not be the most adequate scenario. For example, in medical domain misclassifying an ill patient as a healthy one is usually much more harmful than treating a healthy patient as an ill one and sending him for additional examinations. In finance, investors tend to sell winning stocks more readily than losing stocks in the sense that they realize gains relatively more frequently than losses. The sadness that one experiences in losing the money appears to be greater than the pleasure of gaining the same amount of money. This strong loss aversion was explained and described in the prospect theory by Kahneman and Tversky [14] and applied to finance practice by Shefrin and Statman [25].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_30, © Springer International Publishing Switzerland 2014

In this paper, we want to tackle the cost-sensitive regression methods. We focus on extending the existing *EA* for model tree induction to handle data with asymmetric costs.

1.1 Background

The decision trees [22] are one of the most widely used prediction techniques. Ease of application, fast operation and what may be the most important, effectiveness of decision trees, makes them powerful and popular tool [15]. Regression and model trees [13] may be considered as a variant of decision trees, designed to approximate real-valued functions instead of being used for classification tasks. The main difference between regression tree and model tree is that, in the latter, constant value in the terminal node is replaced by a regression plane. Each leaf of the model tree may hold a linear (or nonlinear) model whose output is the final prediction.

Problem of learning an optimal decision tree is known to be NP-complete. Consequently, classical decision-tree learning algorithms are built with a greedy top-down approach [21] which usually leads to suboptimal solutions. Recently, application of *EA*s [18] to the problem of decision tree induction [2] become increasingly popular alternative. Instead of local search, *EA* performs a global search in the space of candidate solutions. Trees induced with *EA* are usually significantly smaller in comparison to greedy approaches and highly competitive in terms of prediction accuracy [17,7]. On the other hand, the induction of global regression and model trees is much slower [8]. One of the possible solutions to speed up evolutionary approach is a combination of *EA*s with local search techniques, which is known as Memetic Algorithms [12].

Cost-sensitive prediction is the term which encompasses all types of learning where cost is considered [28,10] e.g., costs of tests (attributes), costs of instances, costs of errors. In this paper, we only focus on asymmetric costs, which are associated with different types of prediction errors.

The vast majority of data mining algorithms is applied only to the classification problems [27] while cost-sensitive regression is not really studied outside of statistic field [3]. In induction of cost-sensitive classification trees, three techniques are popular:

- convert classical decision tree into cost-sensitive one, mainly by changing the splitting criteria and/or adopting pruning techniques for incorporating misclassification costs (e.g. [4]);
- application of *EA*s that induce cost-sensitive trees [16];
- application of universal methods like: cost instance-weighting [26] or post-hoc tuning solutions e.g. MetaCost [9].

One of the earliest studies of asymmetric costs in regression was performed by Varian [30]. Author propose *LinEx* loss function which is approximately linear on one side and exponential on the other side as an alternative to popular least squared procedures. Application of different loss functions was later extended

[5] to *LinLin* (asymmetric linear) and QuadQuad (asymmetric quadratic) loss functions. In data mining literature there are only few propositions to handle asymmetric costs e.g. in [6] authors propose a modified back-propagation neutral network that applies *LinLin* cost function.

Recently, post-hoc tuning methods for regression, analogous to ones in cost-sensitive classification, were proposed [3]. Solutions minimize average misprediction cost under an asymmetric cost structure for regular regression models post-hoc by adjusting the prediction by a certain amount. In it's extension [31], application of polynomial functions as model adjustment is proposed to improve the cost-sensitive prediction.

1.2 Motivation

Due, to the lack of cost-sensitive regression solutions in data mining literature, one of the good alternatives are the post-hoc tuning methods [3,31]. However, limitations of such algorithms are obvious as the tuning procedure cannot incorporate cost functions during model learning. In addition, when understanding and interpretation of generated decisions/rules is crucial, such technique cannot be applied.

In this paper, we want to show how to extend existing evolutionary induced model trees to successfully predict under asymmetric losses. In case of evolutionary induced model trees, simple modification of the fitness function, alike for classification trees [17] is not enough, as the linear (or non-linear) models in the leaves are usually not evolved but constructed using standard regression techniques [1,7]. Extensions must also affect the search of cost-sensitive models in the leaves. Full search of regression models is usually difficult for real-life, large datasets due to the huge additional solution space to cover. Therefore, in this paper, we propose two memetic operators that can, together with appropriate fitness function, efficiently convert cost-neutral model trees into cost-sensitive ones.

2 Cost-Sensitive Extensions for Evolutionary Induced Model Trees

In this section we present a combination of evolutionary approaches with local search techniques to achieve a cost-sensitive learner. At first, we briefly describe evolutionary evolved model tree called Global Model Tree (*GMT*) [7]. This evolutionary induced model tree will serve as an example to illustrate the proposed extensions and the fitness function to handle data with asymmetric costs.

2.1 Global Model Tree

GMT follows a typical framework of evolutionary algorithms [18] with an unstructured population and a generational selection. Model trees are represented in their actual form as typical univariate trees. Each test in a non-terminal node concerns only one attribute (nominal or continuous valued). At each leaf a multivariate linear model is constructed using standard regression technique [20] with instances and attributes associated with that node.

Initial individuals are created by applying the classical top-down algorithm [21]. Ranking linear selection [18] is used as a selection mechanism. Additionally, in each iteration a single individual with the highest value of fitness function in current population in copied to the next one *(elitist strategy)*. Several variants of cross-over and mutations were proposed [7,8] that involve:

- exchanging tests, nodes, subtrees and branches between the nodes of two individuals;
- modifications in the tree structure (pruning the internal nodes and expanding the leaves);
- changing tests in internal nodes and extending, simplifying, changing linear regression models in the leaves.

The Bayesian information criterion (BIC) [23] is used as a fitness function and its formula is given by:

$$Fit_{BIC}(T) = -2 * ln(L(T)) + ln(n) * k(T), \tag{1}$$

where $L(T)$ is maximum of likelihood function of the tree T, $k(T)$ is the number of model parameters and n is the number of observations. The log(likelihood) function $L(T)$ is typical for regression models [11] and can be expressed as:

$$ln(L(T)) = -0.5n * [ln(2\pi) + ln(SS_e(T)/n) + 1], \tag{2}$$

where $SS_e(T)$ is the sum of squared residuals on the training data of the tree T. The term $k(T)$ can also be viewed as a penalty for over-parametrization and has to include not only the tree size (calculated as the number of internal nodes) but also the number of attributes that build models in the leaves.

2.2 Cost-Sensitive Extensions

Extending regular regression models to be cost-sensitive requires several steps. At first, appropriate measurement must be defined for assessing the performance of solutions. In our work, we use the average misprediction cost proposed in [3].

Let the dependent variable y be predicted based on a vector of independent variables x. A regression method learns a prediction model, $f : x \to y$ from n training instances. If the function $C(e)$ characterize the cost of a prediction error e then average misprediction cost denoted as Amc can be defined as:

$$Amc = \frac{1}{n}\sum_{1}^{n} C(f(x_i) - y_i). \tag{3}$$

Next, to find cost-sensitive regression models in the tree leaves, we propose BIC extension as fitness function and two local search components that are built into the mutation-like operator.

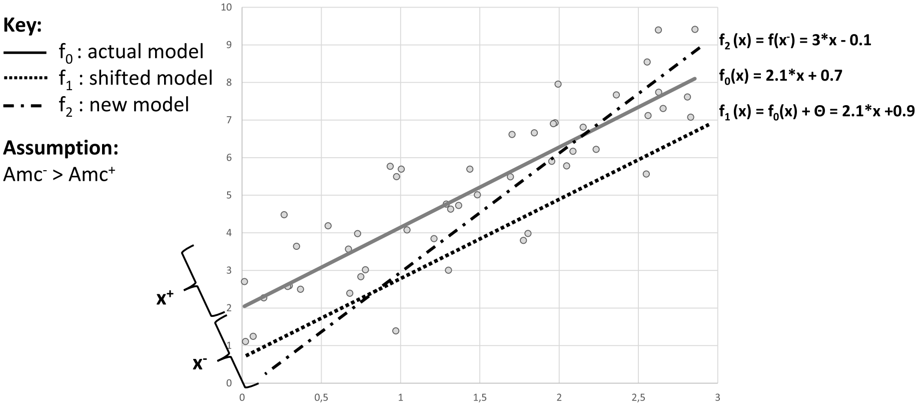


Fig. 1. An example of simple linear regression model $f_0(x)$ changed by cost-sensitive extensions - shift: $f_1(x)$ and new model: $f_2(x)$

Fitness Function. We propose a cost-sensitive BIC to work as a fitness function. We have replaced the squared error loss $SS_e(T)$ from Equation 2 with the average misprediction cost. To remain balance between complexity term $k(T)$ and the cost of the tree, we performed additional experimental research to determine the appropriate value of penalty term, which is now equal $(Q(T) + M(T))$ where $Q(T)$ is the number of internal nodes in model tree T and $M(T)$ is the sum of all attributes in the linear models in the leaves.

Shift Regression Model. The idea of our first mutation variant is similar to the one for cost-sensitive post-hoc tuning method [3]. With the user defined probability, regression model in the leaf is adjusted by a certain amount denoted as θ. Let x^+ represents instances that are over-predicted and x^- instances under-predicted by an actual regression model in the leaf. The costs for over-prediction and under-prediction are equal C^+ and C^-, respectively.

We calculate average misprediction cost separately for x^+ and x^-, denoted as Amc^+ and Amc^- and define the shift θ as:

$$\theta = \begin{cases} -\frac{Amc^+}{C^+} * \delta, \text{ if } Amc^+ > Amc^- \\ \\ \frac{Amc^-}{C^-} * \delta, \text{ if } Amc^+ < Amc^- \end{cases}, \tag{4}$$

where δ is equal:

$$\delta = \frac{Amc^+ - Amc^-}{Amc^+ + Amc^-} rand(0, 1). \tag{5}$$

Main role the parameter δ is to reduce impact of adjustment when Amc on both sides of regression model is similar. Multiplication with a random value from 0 to 1 (denoted as $rand(0, 1)$) extends the number of possible values of θ. Finally, the actual regression model in the leaf is updated by adding the calculated adjustment:

$$f_{new}(x) = f(x) + \theta. \tag{6}$$

It is illustrated in Figure 1 where actual regression model $f_0(x)$ is replaced by the shifted one $f_1(x)$.

New Cost-Sensitive Model. Second variant of mutation replaces actual regression model with a new one that is built on the subset of instances. If, for the actual model in the leaf, the $Amc^+ > Amc^-$ then new cost-neutral regression model is calculated only for over-predicted instances (x^+), otherwise only for under-predicted (x^-). Next, the actual regression model is replaced by the new one:

$$f_{new}(x) = \begin{cases} f(x^+), \text{ if } Amc^+ > Amc^- \\ \\ f(x^-), \text{ if } Amc^+ < Amc^- \end{cases} . \tag{7}$$

In contrast to the first extension, this technique allows finding a completely new model that can decrease Amc for the leaf. Figure 1 illustrates how actual regression model $f_0(x)$ is replaced by the new one denoted as $f_2(x)$, calculated for the x^-.

3 Experiments

We have modified cost-neutral GMT algorithm to show, how proposed extensions handle data with asymmetric costs. In this section we show the performance of $CS - GMT$ (GMT with applied cost-sensitive extensions) on loan charge-off forecasting data. Thanks to the source code of cost-sensitive tuning method and its extensions received from authors [3,31] we are able to compare $CS - GMT$ with post-hoc tuning methods.

3.1 Datasets and Setup

In the paper we used loan charge-off forecasting data from Wharton Research Data Services ($WRDS$, http://wrds-web.wharton.upenn.edu). This data is characterized by asymmetric costs on misprediction errors, because under-prediction of loan charge-off is more costly than over-prediction. If the bank over-predicts its future loan charge-off, the worst what could happen is the reduction of bank's income because there will maintain some extra funds in the loan-loss reserves. The under-prediction means that the bank did not prepare sufficient provisions for its loan losses and has not enough reserves which can cause regulatory problems and significant downturn of its credit rating which is much more dangerous to the bank.

We used the same settings to prepare and test data as in [31], however, more recent data were used. In the experiments, 28 quarters from period $2004 - 2010$ were used with 14 variables related to bank current financial data (described and listed in [3,31]), including loan charge-off, in a particular quarter as the independent variable. The dependent variable is the loan charge-off in the following quarter so the bank can use all useful information while predicting the next quarter loan charge-off.

We generated 27 datasets from 28 quarters because from the last quarter of 2010 only loan charge-off value were used as 2011 data were not available in $WRDS$ yet. For each dataset, prediction model was trained on one quarter and tested on the next one and so on. Therefore, there were 26 training datasets (third quarter of 2010 was used only for testing) and 26 independent testing datasets (first quarter of 2004 was used only for training). In addition, observations with missing values were removed and, to reduce the extent of skewness, the natural logarithm transformation was performed. Average number of instances in each quarter equals 7695 (minimum: 6992 and maximum: 8315). Following [31], we used $LinLin$ cost function and examined cost ratios for under-prediction to over-prediction as follows: $10:1$, $20:1$, $50:1$ and $100:1$. The same three base regression models: standard least-squares linear regression (LR), $M5$ model tree [29] and back-propagation neutral network (NN)[24] were post-hoc tuned for the comparison purpose to $CS-GMT$. Original settings for all tuned methods and $CS-GMT$ solution were applied through all experiments.

3.2 Results

Table 1 summarizes the results of the Amc for three base regression methods tuned by the algorithms described in [3,31] and proposed $CS-GMT$ solution. Each reported quantity is an average value over 26 independent testing datasets (over 200 000 tested instances). The $NONE$ column refers to the results without tuning or cost-sensitive extensions, BSZ refers to the tuning method proposed by Bansal et al. [3] and $LINEAR$ is a linear extension of BSZ algorithm by Zhao et al. [31]. Finally, last column shows the results of $CS-GMT$: proposed cost-sensitive extensions denoted as $CS\ extensions$ applied to GMT.

Results enclosed in Table 1 show that the both post-hoc tuning methods improves the performance of regression model. The extension of BSZ called $LINEAR$, like it was shown in the paper [31], is significantly better than its predecessor. When only post-hoc tuned algorithms are considered, we can observe that the best performance is achieved by NN. However, when we focus on the last column, we see that Amc can be decreased even more. The $CS-GMT$ solution outperforms all tuned base regression models under every cost ratio. Wilcoxon signed rank test for $CS-GMT$ and linearly tuned NN under every cost ratio showed that the differences of Amc between both algorithms are statistically significant ($P\ value < 0.0001$). There is also a significant difference between linearly tuned GMT and $CS-GMT$ which suggests that there is a still significant space for improvement for tuned methods.

The cost reduction between the best out of three linearly tuned algorithms (NN) and $CS-GMT$ is in the range of 7.7% to 9.4% which may be seen by some as not very impressive. However, we must remember that the cost values are on a natural log scale as the values of dependent variable loan charge-off

Table 1. Average misprediction costs for post-hoc tuned base regression algorithms and cost-sensitive extensions for Global Model Tree

Algorithm	Cost ratio	NONE	BSZ	LINEAR	CS extensions
LR	10	7.41	3.78	3.81	-
M5	10	7.29	4.16	3.88	-
NN	10	8.16	3.69	3.57	-
GMT	10	7.07	3.81	3.65	3.29
LR	20	14.06	4.84	4.42	-
M5	20	13.78	5.47	4.86	-
NN	20	15.60	4.62	4.26	-
GMT	20	13.66	5.07	4.27	3.85
LR	50	34.02	6.24	5.23	-
M5	50	33.23	6.03	6.44	-
NN	50	37.92	5.69	5.09	-
GMT	50	32.96	6.40	6.11	4.66
LR	100	67.27	7.06	5.85	-
M5	100	65.66	7.24	7.94	-
NN	100	75.12	6.50	5.80	-
GMT	100	64.15	7.03	5.98	5.27

were transformed by the natural logarithm. Therefore, the real cost reduction on the original scale is in range 24.2% to 40.7% and therefore, can be attractive in bank loan charge-off forecasting problem.

Application of proposed cost-sensitive extensions does not only significantly reduce *Amc*. The important benefit of proposed extensions, in context of *GMT* is that the whole tree: test in internal nodes and models in the leaves, fits to the analyzed, cost-sensitive problem. Therefore, decisions from $CS-GMT$ are much easier to interpret. Identifying patterns and finding explanations for predictions may be difficult for tuned regression models because all rules obtained by the phase of learning were cost-neutral.

4 Conclusion and Future Works

In this paper, we propose extensions for evolutionary induced model trees to achieve cost-sensitive learners. We present a cost-sensitive *BIC* to work as the fitness function and allows the algorithm to minimize the average misprediction cost. Two specialized memetic operators that search for cost-sensitive regression models in the tree leaves are also proposed. Those local optimizations of the regression models are simple, complementary and easy to apply.

Experiments performed of 27 real-life datasets show that there is a significant difference between post-hoc tuned methods and solutions that explicitly incorporate cost functions during model building - like proposed $CS - GMT$. The real, average cost reduction between the best out of 4 tuned algorithms (linearly tuned *NN*) and proposed $CS-GMT$ is over one third. In addition, as generated decisions and models from $CS - GMT$ take into account costs during learning phase, they can be used to learn and understand underlying processes from the data.

There are a number of promising directions for future research. In particular, we should consider testing different cost functions and handling different types of costs like e.g. cost of attributes. Application of the proposed extensions to other EA solutions that construct models with standard regression techniques and testing other forecasting problems requires further extensive research, which we leave for the future.

Acknowledgments. This work was supported by the grant W/WI/1/2013 from Bialystok University of Technology. The authors thank to Huimin Zhao, Atish P. Sinha and Gaurav Bansel for providing us with implementation of their tuning methods.

References

1. Barros, R.C., Ruiz, D.D., Basgalupp, M.: Evolutionary model trees for handling continuous classes in machine learning. Information Sciences 181, 954–971 (2011)
2. Barros, R.C., Basgalupp, M.P., Carvalho, A.C., Freitas, A.A.: A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems Man and Cybernetics, Part C 42(3), 291–312 (2012)
3. Bansal, G., Sinha, A.P., Zhao, H.: Tuning data mining methods for cost-sensitive regression: a study in loan charge-off forecasting. Journal of Management Information Systems 25(3), 317–338 (2008)
4. Bradford, J., Kunz, C., Kohavi, R., Brunk, C., Brodley, C.E.: Pruning decision trees with misclassification costs. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 131–136. Springer, Heidelberg (1998)
5. Cain, M., Janssen, C.: Real estate price prediction under asymmetric loss. Annals of the Institute of Statistical Mathematics 47(3), 401–414 (1995)
6. Crone, S.F., Lessmann, S., Stahlbock, R.: Utility based data mining for time series analysis: Cost-sensitive learning for neural network predictors. In: Proc. of 1st UDBM, Chicago, IL, pp. 59–68 (2005)
7. Czajkowski, M., Kretowski, M.: An Evolutionary Algorithm for Global Induction of Regression Trees with Multivariate Linear Models. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) ISMIS 2011. LNCS, vol. 6804, pp. 230–239. Springer, Heidelberg (2011)
8. Czajkowski, M., Kretowski, M.: Does Memetic Approach Improve Global Induction of Regression and Model Trees? In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) SIDE 2012 and EC 2012. LNCS, vol. 7269, pp. 174–181. Springer, Heidelberg (2012)
9. Domingos, P.: MetaCost: A general method for making classifiers cost-sensitive. In: Proc. of KDD 1999, pp. 155–164. ACM Press (1999)
10. Elkan, C.: The Foundations of Cost-Sensitive Learning. In: Proc. of IJCAI, pp. 973–978 (2001)
11. Gagne, P., Dayton, C.M.: Best Regression Model Using Information Criteria. Journal of Modern Applied Statistical Methods 1, 479–488 (2002)

12. Gendreau, M., Potvin, J.Y.: Handbook of Metaheuristics. International Series in Operations Research & Management Science, vol. 146 (2010)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd edn. Springer (2009)
14. Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decisions under Risk. Econometrica 47(2), 263–292 (1979)
15. Kotsiantis, S.B.: Decision trees: a recent overview. Artificial Intelligence Review, 1–23 (2011)
16. Krętowski, M., Grześ, M.: Evolutionary Induction of Cost-Sensitive Decision Trees. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 121–126. Springer, Heidelberg (2006)
17. Kretowski, M., Grześ, M.: Evolutionary Induction of Mixed Decision Trees. International Journal of Data Warehousing and Mining 3(4), 68–82 (2007)
18. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 3rd edn. Springer (1996)
19. Murthy, S.: Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery 2, 345–389 (1998)
20. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C. Cambridge University Press (1988)
21. Rokach, L., Maimon, O.Z.: Top-down induction of decision trees classifiers - A survey. IEEE Transactions on Systems Man and Cybernetics, Part C 35(4), 476–487 (2005)
22. Rokach, L., Maimon, O.Z.: Data mining with decision trees: theory and application. Machine Perception Arfitical Intelligence 69 (2008)
23. Schwarz, G.: Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464 (1978)
24. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing, pp. 318–362. MIT Press, Cambridge (1986)
25. Shefrin, H., Statman, M.: The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. Journal of Finance 40, 777–790 (1985)
26. Ting, K.: An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14(3), 659–665 (2002)
27. Torgo, L., Ribeiro, R.: Utility-based regression. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 597–604. Springer, Heidelberg (2007)
28. Turney, P.: Types of cost in inductive concept learning. In: Proc. of ICML 2000 Workshop on Cost-Sensitive Learning, Stanford, CA (2000)
29. Quinlan, J.: Learning with Continuous Classes. In: Proc. of AI 1992, pp. 343–348. World Scientific (1992)
30. Varian, H.R.: A Bayesian Approach to Real Estate Assessment. In: Fienberg, S.E., Zellner, A. (eds.) Studies in Bayesian Econometrics and Statistics: In honor of L.J. Savage, North-Holland, Amsterdam, pp. 195–208 (1974)
31. Zhao, H., Sinha, A.P., Bansal, G.: An extended tuning method for cost-sensitive regression and forecasting. Decision Support Systems 51, 372–383 (2011)

Optical Music Recognition as the Case of Imbalanced Pattern Recognition: A Study of Complex Classifiers

Agnieszka Jastrzebska[1] and Wojciech Lesinski[2]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland

[2] Faculty of Mathematics and Computer Science, University of Bialystok,
ul. Sosnowa 64, 15-887 Bialystok, Poland

Abstract. The article is focused on a particular aspect of classification, namely the imbalance of recognized classes. Imbalanced data adversely affects the recognition ability and requires proper classifier's construction. The aim of presented study is to explore the capabilities of classifier combining methods with such raised problem. In this paper authors discuss results of experiment of imbalanced data recognition on the case study of music notation symbols. Applied classification methods include: simple voting method, bagging and random forest.

Keywords: music recognition, ensemble classifiers, imbalanced data.

1 Introduction

The task of image recognition and classification has been known and addressed in the literature since many years. One of the most significant issues negatively influencing the results of visual data mining is poor balance of classes. Several issues concerning imbalanced data have been discussed in [5] and [3]. An important research area, where this problem has not been yet successfully overcome is musical notation symbols classification.

The issue of optical character recognition is very important in modern computer science. Recognition of graphic characters, such as numbers and letters, greatly simplifies digitization of documents of any type, which is especially important during the transition to electronic mass communication. Even today one can submit in this form an annual tax return, or settle many other official issues. Optical Character Recognition (OCR) facilitates this procedure, and greatly accelerates the process of transition into the digital information era. Another important aspect of automatic image recognition and understanding is related to our efforts to improve the quality of life of blind and visually impaired. Programs that recognize printed text, but also handwritten digits are in common use. Their effectiveness is high, but searching for more recent and better solutions is always possible. Unfortunately, in other areas, yet much remains to

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_31, © Springer International Publishing Switzerland 2014

be done. Fields such as handwriting recognition, face identification, signature verification and recognition, and many others are still waiting for the efficiency improvement of existing algorithms. An important area of visual data mining is the optical recognition of musical notation.

Automatic recognition and classification of music notation may have many applications. This is primarily a musical scores backup. Electronic processing of acquired information could be another application. With electronic record of music notation we can attempt to computerize musical synthesis, we can also, by using the voice synthesizer, read this music score for the need of blind and visually impaired. Electronic music notation could also be used to verify the performances correctness of the musical composition, and to detect potential plagiarism. These applications lead to the conclusion that the optical recognition of music notation is an interesting and worthy research topic

General methodology of optical music recognition has been already researched and described in [7] and [11]. We would like to highlight, that studied problem of imbalance of classes is an original contribution to the field of music symbols classification. The aim of our study is to investigate how well-known classification tools deal with imbalanced data. In this paper presented are complex classifiers only. The research is based on actual opuses. Applied classification algorithms have been implemented in C++. Developed program works with both high and low-resolution images of musical symbols.

The paper is organized as follows. Section 2 lists the basic information about the classification and used classifiers. In Section 3 the learning and testing sets are outlined. Section 4 describes empirical tests.

2 Preliminaries

2.1 Classifiers Conjunction

In this paper we apply various complex classifiers. They join computational capabilities of single classifiers and allow to build diverse models. In the case of conjunction methods, classifier is created with a number of other classifiers. Classifiers, which we use for connecting, we can call "weak classifiers". Depending on the purpose, we may compose a model consisting of various single classifiers, but we may also manipulate with distinct parameters. There are also different ways of model construction.

Classifier ensembles are typically applied, when there is a risk that a single classifier would perform inadequately. Generally, there are four different strategies of ensemble classifier construction. Main differences are in: methods of output averaging or weighting, selection of base classifiers, selection of best features subset, selection of different data subsets.

In ensemble learning tested sample x is recognized by all used classifiers, next results are compared to adjust only one system response. Usually, the process of joining results is nothing else, but sum of answers with set weights.

2.2 Simple Voting

Simple voting is one of the simplest conjunction methods. We can use any component classifiers in this method. Classifiers can be already trained, or they can be in the phase of training. The way of training is also not imposed. The only condition of start-up of this algorithm is having classifiers, which are statistically independent from each other. The sample $x \in X$ is tested by every weak classifier, then an answer is counted as a sum. The class which is indicated by most of classifiers, is chosen as the right one.

2.3 Bagging

Bagging, a name derived from "bootstrap aggregation", devised by Breiman [1], is one of the most intuitive and simplest ensemble algorithm providing good performance. It uses multiple versions of a training set, each created by drawing $n = N$ (where N is a number of elements of original training set) samples from training set D with replacement. Each of bootstrap data sets is used to train a different component classifier and the final classification decision is based on the vote of each component classifier. Traditionally the component classifiers are of the same general form - for example, all Hidden Markov models, or all neural networks, or all decisions trees - merely the final parameter values differ among them due to their different sets of training patterns.

2.4 Random Forest

Random forest is a relatively new classifier proposed by Breiman in [2]. The method combines "bagging" idea [1] and the random selection of features in order to construct a collection of decision trees with controlled variation.

Random forest is composed of some number of decision trees. Each tree is built as follow:

- Let the number of training objects be N, and the number of features in features vector be M.
- Training set for each tree is built by choosing N times with replacement from all N available training objects.
- Number $m << M$ is an amount of features on which to base the decision at that node. This features is randomly chosen for each node.
- Each tree is built to the largest extent possible. There is no pruning.

Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). A random features selection used in the subsequent divisions of a single tree prevent over-fitting to training data. No pruning allows the use of the ID3 algorithm proposed by Quinlan in [12].

3 Data Set

The recognized set of music notation symbols had about 27.000 objects in 20 classes. There were 12 numerous classes and each of them had about 2.000 representatives. Cardinality of the other eight classes was much lower and various in each of them. Part of the examined symbols was cut manually from chosen Fryderyk Chopin's compositions. Other part of the symbols' library comes from research projects [13] and [14].

3.1 Recognized Symbols

As a regular class considered are: flat, sharp, natural, G clef, F clef, forte,piano, mezzo forte, quarter rest, eight rest, sixteenth rest and flagged stem. To irregular classes breve note, accent, crescendo, diminuendo, tie, fermata, C clef and thirty-second rest were included.

4 Results

The experiment was divided into two parts. In the first one only elements belonging to the regular class were being recognized. It allowed to determine the appropriate structure of classifiers. In this case the assessment of classifiers was basing on the:

- accuracy calculated by the equation:

$$acc = \frac{number\ of\ well\ recognized\ objects}{number\ of\ all\ objects} \tag{1}$$

- classifier's error:

$$err = \frac{number\ of\ objects\ recognized\ incorrectly}{number\ of\ all\ objects} \tag{2}$$

In the second stage irregular classes were added to the previously recognized classes. At this point, attention was paid to changes in the efficiency of recognition and recognition accuracy of particular irregular classes. Apart from *acc* and *err*, measure showing the influence of classes counting less elements should be used for classifiers assessment. Some of these measurements were described in paper [5]. Of these, we will use coefficients $TPrate$ and $FNrate$ showing the effectiveness of the selected class.

$$TPrate = \frac{TP}{TP+FN} = \frac{number\ of\ good\ classification\ in\ a\ given\ class}{the\ number\ of\ all\ objects\ from\ a\ given\ class} \tag{3}$$

$$FNrate = \frac{FN}{TP+FN} = \frac{number\ of\ good\ classification\ in\ a\ given\ class}{the\ number\ of\ all\ objects\ from\ a\ given\ class} \tag{4}$$

4.1 Features Vector

Classification was done on features characterizing every symbol. Predictive features were defined based on theoretical study in the area of image recognition and on the experience of authors. Following features were used in the experiment:

Projections Horizontal and vertical projections in a rectangle were taken. Horizontal projection is defined for every row of pixels of the rectangle. For a given row, the value of the projection is equal to the number of black pixels in this row. By analogy, vertical projection is defined in columns of the rectangle. For both projections, the maximum value, position of the maximum value, the average value and the support (number of nonzero values) were included in the feature's vector.

Transitions Like in the case of projections, horizontal and vertical transitions in a rectangle were taken. Horizontal transitions are defined for every row of pixels of the rectangle. For a given row, the value of the transition is equal to he number of pairs of consecutive white and black pixels in this row. By analogy, vertical transitions are defined in columns of the rectangle. Transitions reflect shape complexity of the image. Maximal values of transitions in both horizontal and vertical directions were included in the features' vector.

Margins Left, top, right and bottom margins in a rectangle were taken. Left margin is defined for every row of pixels of the rectangle. For a given row, the value of the left margin is equal to the number of white pixels from the left edge of the rectangle right to the first black one. The value of the right margin is equal to the number of white pixels from the right edge of the rectangle left to the first black one. Top ad bottom margins are defined analogously. These features show the symbol's position in the image. We used maximum value of all margins in features vector.

Moments are used in different fields, e.g. in physics (e.g. mass, center of mass, moment of inertia), in probability (e.g. mean value, variance). In image processing, computer vision and related fields, moments are certain particular weighted averages of the image pixels' intensities. Also, functions of moments are often utilized in order to have some attractive property or interpretation. Image moments are useful to describe objects after segmentation. Simple properties of the image which are found via image moments including area (or total intensity), its centroid and information about its orientation.

Directions for given pixel it is the longest segment of black pixels in given directions (usually horizontal, vertical, left and right diagonal directions are considered) which include given black pixel. Directions of 22.5, 67.5, 112.5 and 157.5 degrees were considered too.

Derivatives Another group of features are derivatives of the afore-mentioned vector features. Derivative is defined as the vector of differences between subsequent elements of a given vector. We used derivatives of the projection, transition and margins. From derivatives, we obtained additional information, which we were unable to gain from the origin vectors.

Others features such as field and circuit of the symbol, Euler number, eccentricity and quarter.

4.2 Regular Classes

Simple Voting. A simple voting method is the simplest complex classifier. Basing on previous studies, in this algorithm we have combined following classifiers: kNN ($k = 1$), Mahalanobis minimum distance and decision tree. In the case of a draw object was assigned to the class, which was indicated by decision tree. The effectiveness of this method was growing with the enlargement of the training set. Comparing to the simple classifiers, it was more effective for small number of training data, while for larger data sets these results were similar. Plot illustrating results for this classifier is shown in Figure 1. Accurate results for particular classes for learning sets composed of 400 elements are shown in Table 3.

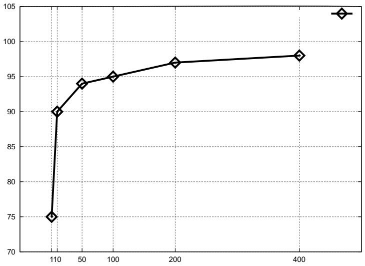


Fig. 1. Simple voting accuracy (in %) collated with the number of representatives of each class in the training set for regular classes

Bagging. Another complex classifier used, which we have used was bagging. In the performed tests we have combined: kNN ($k = 1$) and decision tree. In each run built were 10 classifiers apiece from given type. Also in this case, classifier's performance increased with enlargement of the training sets. The overall efficiency was higher, than in the case of simple classifiers. This difference was particularly noticeable for tests carried out on the small learning sets. When the training set counted 400 elements for each class, this difference was less than 0.5%. The function of the recognition efficiency depending on the size of the training set is shown in Figure 2. The exact results in particular classes for bagging based on decision trees for learning sets counting 400 elements are shown in Table 3.

Random Forest. In order to apply this classifier, firstly we had to determine optimal model parameters. Of interest were:

- number of trees in the forest,
- number of features randomly chosen to divide a single node.

A study on the number of trees in the forest was carried out as the first. The training sets counting 100 elements in each class were used for building each tree. In order to divide a single node, five available features were randomized. Forests

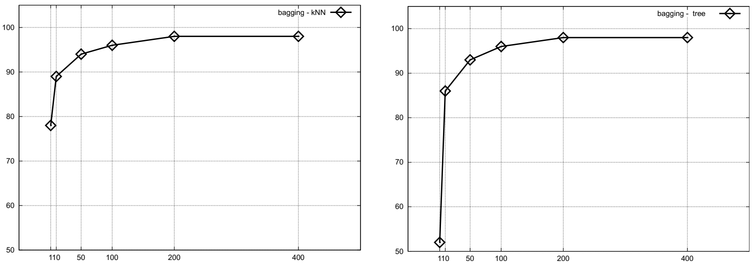


Fig. 2. Bagging's accuracy (in %) collated with the number of representatives of each class in the training set for regular classes

consisting of one and three trees only obtained worse results than a typical decision tree running on the same learning set. With the increase of the number of trees, recognition rate grew and became better than in case of usual tree. For 20 trees efficiency reached the level of 98% and ceased to grow significantly. All results are in Table 1. The next step was to determine the optimal number of features randomized during the construction of a single node. Forest counting 10 trees was used for the tests. Training data set consisted of 100 elements in every class. The behavior of the classifier was examined for 1, 2, 3, 4, 5, 8, 10, 15 and 20 randomized features. The results are shown in Table 2. The poorest efficiency characterizes the forest, which was randomly selecting only one feature. This outcome is coherent with our expectations. Completely random selection of features and total rejection of any other measures of division result in poor accuracy. As the number of randomized features becomes greater, the efficiency increases. It should be noticed, that if we increase the number of features over 5, the increase in accuracy becomes no longer significant.

Table 1. Influence of the number of trees and the efficacy of the random forest

number of trees	effectiveness (in %)
1	92
3	93
5	95
10	96
20	97
50	97
100	97

The studies were also conducted to find out, how the training set's size affects the efficiency of the random forest method. During this research forest was built with 20 trees, while for the division in a single node 5 available features were randomized. As in previous cases, the effectiveness rose with increasing number of learning data. For sets counting 400 elements in each class, it amounted to over 98% and it was one of the best results. All results are shown in Figure 3.

Table 2. Influence of the number of random features on random forest effectiveness

number of random features	effectiveness (in %)
1	87
2	90
3	92
4	95
5	97
10	97
20	97

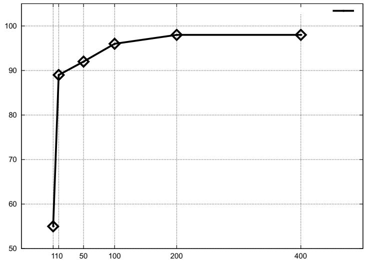


Fig. 3. Random forest's accuracy (in %) collated with the number of representatives of each class in the training set for regular classes

The exact results obtained in particular classes by the random forest for learning sets of up to the 400 elements are shown in Table 3.

4.3 Irregular Classes

In the following step we have added to the training set classes determining the imbalance. In this case, we would like to focus on the effectiveness of recognition within irregular classes. Regular classes had 400 representatives in the training set. The number of elements in irregular classes is presented in Table 4.

Simple Voting. In this method, the following classifiers were combined: kNN ($k = 1$), Mahalanobis distance and decision tree. In case of a draw symbol was classified to the class, which was pointed out by the k-Nearest Neighbors classifier. This algorithm transferred properties of its component classifiers. There was a slight decrease in global performance in comparison to the recognition of only regular classes. Irregular classes were classified less efficiently. The accuracy of rare symbols recognition depends on the cardinality of corresponding training set. The smallest is class representation in data, the poorest are classification results.

The overall effectiveness of the simple voting method was $acc = 98.05\%$, and the error coefficient $err = 1.95\%$. Detailed results for irregular classes are given in Table 5.

Table 3. The effectiveness (in %) of recognition in regular classes

classifier/single class training set size	1	10	50	100	200	400
simple voting method	75	90	94	95	97	98
bagging (tree)	53	86	93	96	98	98
random forest	56	89	92	96	98	98

Table 4. Learning and testing sets for irregular classes

class	learning set	testing set
accent	30	65
breve	1	2
crescendo	55	100
diminuendo	51	97
fermata	35	46
clef C	100	178
tie	100	155
thirty-second rest	20	35

Table 5. Effectiveness of irregular classes recognition (in %)

class	Simple voting		Bagging		Random forest	
	TPrate	FNrate	TPrate	FNrate	TPrate	FNrate
accent	94	6	92	8	94	6
breve	0	100	0	100	0	100
crescendo	90	10	92	8	92	8
diminuendo	91	9	91	9	91	9
fermata	87	13	91	9	93	7
clef C	99	1	99	1	99	1
tie	95	5	97	3	98	2
thirty-second rest	86	14	86	14	91	9

Bagging. Bagging was another complex classifier used in the tests. In the performed experiment we have combined: kNN ($k = 1$) and decision tree. In each run used were 10 classifiers apiece from given type. Rare symbols were identified slightly better than in the case of the simple component classifiers. Bagging based on decision trees gave slightly better results, and we will base our further discussions on it. Its total efficacy was $acc = 98.15\%$, and the error coefficient $err = 1.85\%$. Detailed results for irregular classes are shown in Table 5.

Random Forest. The last of examined methods was the random forest. In the experiment forest was built with 10 trees. To divide a single node 5 available features were randomly chosen. The global effectiveness of the classifier on the data set with irregular classes dropped only by one per cent. As in previous methods symbols from the irregular classes were classified worse than symbols from regular classes. Nevertheless, random forest achieved the best performance in classifying irregular classes.

The overall efficacy of the random forest was $acc = 98.21\%$, and the error coefficient $err = 1.81\%$. Detailed results for irregular classes are in Table 5.

5 Conclusion

In the paper discussed was the problem of imbalanced image recognition on the example of music notation symbols recognition. Authors present results of classification experiments performed with complex classifiers on a dataset consisting of 27 000 elements of 20 classes. Applied were: simple voting, bagging and random forest with experimentally validated parameters and properties. We have performed experiments firstly on regular (balanced) classes and secondly on full, imbalanced data set. We have analyzed and compared results obtained in both cases. Studied methods of combining single classifiers improve the performance of classification.

In summary, 20 classes of musical notation symbols were classified. 12 of them have been considered as regular classes, the other as irregular classes. The recognition effectiveness of regular classes was very satisfying. Among the discussed classifiers the best efficacy had random forest and bagging based on kNN. Unfortunately, kNN classifier if of high computational complexity, what prolongs the duration of action. Therefore, every method, which applies kNN is difficult to use in practice. Random forest, based exclusively on decision trees in this respect is much better. A simple voting method achieved results slightly lower than the random forest or bagging.

In our current research we have investigated classification schemes based on classical sets theory. In such case belongingness is crisp. An element either belongs to given class or not. There are fuzzy generalizations of classical approach to classification. In the next step of our research we will take a closer look at other information representation models and we will investigate their suitability to optical music recognition. We would like to verify, if other approaches, especially ones involving bipolarity (c.f. [6], [8] and [9]) may enhance classification results with imbalanced data.

Acknowledgment. The research is supported by the National Science Center, grant No 2011/01/B/ST6/06478, decision no DEC-2011/01/B/ST6/06478.

References

1. Breiman, L.: Bagging predictors. Machine Learning 26(2), 123–140 (1996)
2. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
3. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. Technical report, Dept. of Statistics, U.C. Berkeley (2004)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, Inc., New York (2001)
5. Garcia, V., Sanchez, J.S., Mollineda, R.A., Alejo, R., Sotoca, J.M.: The class imbalance problem in pattern recognition and learning. In: II Congreso Espanol de Informatica, pp. 283–291 (2007)

6. Homenda, W.: Balanced Fuzzy Sets. Information Sciences 176, 2467–2506 (2006)
7. Homenda, W.: Optical Music Recognition: the Case Study of Pattern Recognition. In: Computer Recognition Systems, pp. 835–842. Springer (2005)
8. Homenda, W., Pedrycz, W.: Balanced Fuzzy Gates. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 107–116. Springer, Heidelberg (2006)
9. Homenda, W., Pedrycz, W.: Symmetrization of Fuzzy Operators: Notes on Data Aggregation. In: Halgamuge, S.K., Wang, L. (eds.) Computational Intelligence for Modelling and Prediction. SCI, vol. 2, pp. 1–18. Springer, Heidelberg (2005)
10. Kuncheva, L.I.: Combining Pattern Classifiers. Methods and Algorithms. John Wiley & Sons (2004)
11. Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A.R.S., Guedes, C., Cardoso, J.S.: Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval 1, 173–190 (2012)
12. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1, 81–106 (1986)
13. Breaking accessibility barriers in information society. Braille Score - design and implementation of a computer program for processing music information for blind people, the research project no N R02 0019 06/2009 supported by by The National Center for Research and Development, Poland (2009-2012)
14. Cognitive maps with imperfect information as a tool of automatic data understanding. Ideas, methods, applications, the research project no 2011/01/B/ST6/06478 supported by the National Science Center, Poland (2011-2014)

Recognition of Vehicle Motion Patterns in Video Sequences

Bartosz Buczek and Urszula Markowska-Kaczmar

Wroclaw University of Technology, Wyb.Wyspianskiego, 50-370 Wroclaw, Poland
urszula.markowska-kaczmar@pwr.wroc.pl

Abstract. In the paper a vision system capable to extract and to recognize vehicle motion patterns in a video sequence is described. The video sequence is an isometric view of a crossroad. There is no additional knowledge about the scene. The novelty of the proposed approach lies mainly in the vehicle tracking method and the way of motion pattern representation. The recognition method is based on the Hidden Markov Model.

Keywords: vehicle tracking, vehicle motion recognition, Hidden Markov Model.

1 Introduction

A phenomenal increase in computer power over last years has caused automatization of many processes that in the past were made manually by human with less precision. This increase enables also a progress in machine vision systems which receive, analyze and interpret an image of a real scene. During the last two decades, machine vision has been applied to a variety of challenging tasks in the manufacturing process, in medical images recognition or in surveillance systems for detecting abnormal activities. Now the interpretation of a behaviour pattern in a video sequence has become a very important topic in the vision system domain.

In our study we focus on the system capable to recognize four car motion patterns on crossroads. We assumed that video sequence is taken from a static camera representing an isometric view of the crossroad. This assumption makes the tracking problem harder than in the case of the top view, because of merging objects in the frame. Additionally, in contrary to the paper [1] we do not use any knowledge about scene, i.e. information about lanes or the location where the cars can arrive was unknown. Such a system can be applied to automatically change the time of cross lights in reference to the traffic or to find abnormalities in the car motion.

The contribution of the paper lies in the new representation of the object trajectory, the tracking algorithm and details of the recognition method.

The paper consists of five sections. Section 2 presents a short survey of related projects. Section 3 presents the details of the system. The section 4 describes

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_32, © Springer International Publishing Switzerland 2014

experiments and their results. Finally, in the conclusions we summarize our research showing advantages and shortcomings of the approach. The perspective for the future improvement of the method is also proposed.

2 Related Projects

Studies on movement analysis represent two groups - tracking based approaches and nontracking approaches. The first ones use trajectories to model a scene. An object trajectory is defined as a set of points representing the ordered observations of the location of a moving object made at different points in time. In many projects tracking is performed in reference to single objects and pattern behaviour recognition pertains to the particular motion of a single object [1]. The mentioned approach is very close to our project. In contrary to the work [2] the authors did not stop their project on object tracking but they classified trajectories in reference to a performed manoeuvre. They processed an input video sequence representing isometric view on crossroad with a huge road traffic. To detect anomalies and predict behaviors statistical methods were used.

To recognize a class of movement patterns a variety of classification methods were applied: SVM [3], neural networks which are very popular in human motion (gesture) recognition [4], but the most popular approach seems to be using Hidden Markov Model (HMM) [5].

3 Overview of the System

The aim of our system was to recognize in the video sequence one of four vehicle motion patterns on the crossroad based on an isometric view of a crossroad. Recognized patterns are: *drive straight*, *turn left*, *turn right* and *turn back*.

It is worth underlying that sequences were recorded using a static and fixed camera, which lets to use a simpler method to foreground extraction. Moreover, the sequences were recorded in natural lighting during the daytime.

The whole video processing schema is illustrated in Figure 1. *The input* is a frame which fulfills all the above mentioned requirements. As *an output* the label of the recognized class of vehicle motion pattern is given. The whole video processing can be divided into four continuously performed stages: preprocessing, blobs extraction, tracking and classification of gathered trajectories. Details of these steps are described below.

3.1 Preprocessing

The input: a raw frame of a video sequence. *The output*: a frame with a reduced noise and extracted those parts of the frame which correspond to moving vehicles. In the first step the input frame is transformed to a gray scale in order to reduce computational complexity of processing. In the next step, background subtraction is performed. The aim of this step is extraction of foreground pixels which correspond to moving regions on a frame. We focused on sequences recorded by a static camera, so it was possible to use a method with dynamic background modelling. For all frames a background model is updated using the equation (1):

$$\beta^{'} = (1 - \alpha) \cdot \beta + \alpha \cdot I, \tag{1}$$

where: $\beta^{'}$ corresponds to updated background model, α is called *adaptation coefficient*, I is the currently processed frame. Foreground pixels are identified by subtracting the background model $\beta^{'}$ from the processed frame I. In the next step, thresholding is performed in order to get a binary image with non-zero pixels related to foreground objects. During background subtraction, some pixels can be faulty recognized as a foreground. It can be caused by noises in the input frame. To reduce that regions, the median blur is performed at the end of the preprocessing stage.

3.2 Blobs Extraction

The input: the preprocessed frame with designated regions corresponding to moving objects. *The output*: a frame with locations of separated regions corresponding to moving objects. In order to obtain information about their location

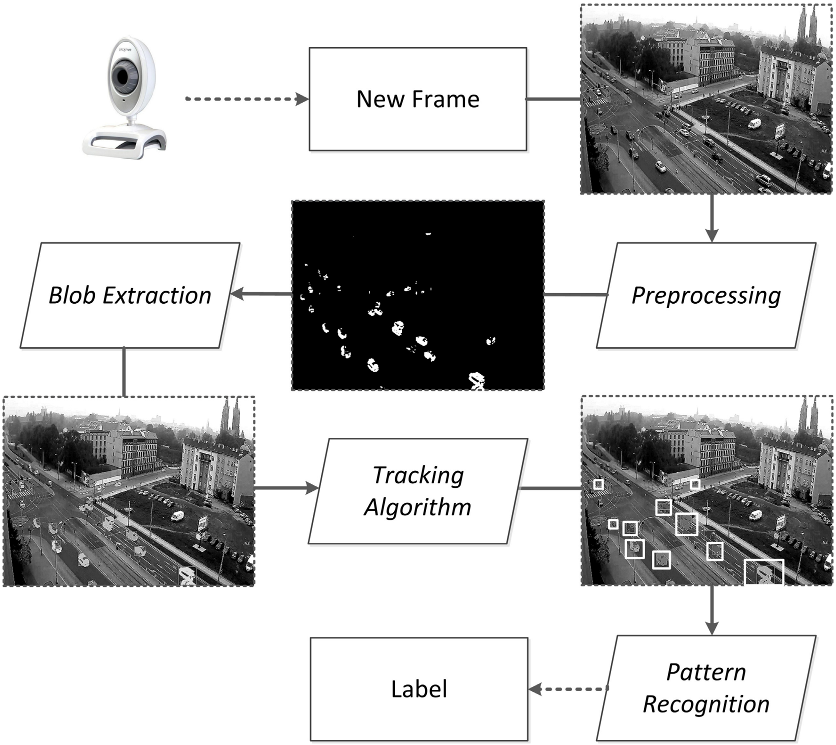


Fig. 1. Processing schema of the whole system

and to enable a separation of regions of moving objects a blob (binary large objects) extraction is performed. Using information about location of particular objects on a scene, we can easily appeal to the original frame which is used in the tracking algorithm.

Binary large objects are extracted on the basis of clustering foreground pixels using labelling algorithm ([6]). Both, a set of binary objects and centroid coordinates representing blobs are returned as an outcome of this stage. Blobs are surrounded by bounding boxes.

3.3 Vehicle Tracking Algorithm

The input: the video sequence with locations of extracted regions corresponding to the moving objects. *The output*: a set of trajectories of moving objects. Tracking of multiple vehicles is the most complex part of the whole system because of two assumptions referring to our system: tracking of vehicles is performed without any information about a structure of a scene and the scene contains an isometric view on crossroads. The last assumption causes many difficulties during tracking due to merging two or more objects into one, large object (blob) extracted in the previous stage. Such situation would not be observed when sequences are recorded using view from the top.

Let o_t^l represents a set of points belonging to the blob (object) l on the frame t. A set of objects gathered in the previous stage creates an input to the tracking algorithm. For each frame t the tracking algorithm strives to find corresponding objects from the previous frame $t-1$ and next, centroid coordinates of all objects are recorded. When an object disappears from a scene, a sequence of particular positions of its centroid on all frames creates a vehicle trajectory which is returned as an outcome of the algorithm.

The following cases were considered during designing the tracking algorithm:

- searching for continuation of an object, i.e. for each o_{t-1}^k from the frame $t-1$ a corresponding object o_t^l is assigned from the frame t,
- automatic detection of objects disappearing from a scene; in this case the object is no longer tracked,
- splitting one object into two or more objects on the next frame; this case refers to the situation where the object o_{t-1}^k from the frame $t-1$ corresponds to two or more objects $o_t^{l_n}$ in the frame t,
- merging two or more objects into one object on the next frame; this means that two objects $o_{t-1}^{k_n}$ and $o_{t-1}^{k_v}$ from the frame $t-1$ are jointed and on the frame t they correspond to one object o_t^l.

In an optimistic case, when all objects can be easily separated, recognition of continuing object movement can be implemented as finding the nearest blob for a particular blob on the current frame but in reality additional information is necessary, such as appearance of objects and the location prediction. At the beginning for object o_{t-1}^l in the frame $t-1$ a subset $Z_k = \{o_t^1, o_t^2, \ldots, o_t^d\}$ of all objects within r distance from the predicted centroid of object o_{t-1}^l is selected.

These objects are called candidates. Then, from Z_k the nearest object o_t^l is selected. The following conditions should be satisfied in this case:

- the recognized objects (blobs) should be greater than an assumed threshold,
- the distance between o_t^l and o_{t-1}^k is the smallest for all objects from the frame $t-1$,
- the difference between average color of objects o_{t-1}^k and o_t^l is less than assumed threshold value.

During tracking the position of the object centroid is predicted using alpha-beta filter [7]. Automatic detection of objects disappearing from a scene is made by counting frames in which continuation of object cannot be found. When the counter exceeds a threshold value, the object is removed from a tracking list and then its trajectory is forwarded to the next stage.

The next problem is splitting object into two or more objects on the next frame. This situation occurs when object is partially obscured by a static object or when foreground pixels are not correctly identified. In the first case for the object o_{t-1}^k the nearest object o_t^l is assigned and then other objects $o_t^i \in Z_k$ are merged if the distance $d_{bb}(o_{t-1}^k, o_t^i)$ is less than a threshold value. The distance d_{bb} is defined as follows:

$$d_{bb}(o_1, o_2) = \begin{cases} 0, & P_1 \cap P_2 \neq \emptyset \\ min(d(A,B)), & A \in P_1, \quad B \in P_2 \end{cases} \tag{2}$$

In eq. 2. the symbol $d(A, B)$ means the Euclidean distance between points A and B, P_1 and P_2 are bounding boxes assigned to objects o_1 and o_2. The distance between two objects is equal to 0 if they are overlapped.

Splitting is also applied when on a scene a new object o_t^p appears such that in its neighbourhood exists other object o_t^l which has rapidly changed its size. Then the trajectory from o_t^l translated by the distance between centroids of o_t^p and o_t^l is assigned to o_t^p .

Another difficult situation that must be considered in the tracking algorithm is merging two objects $o_{t-1}^{k_1}$ and $o_{t-1}^{k_2}$ into one object o_t^l on the next frame t. In this case, when as a continuation for the object $o_{t-1}^{k_1}$ the object o_t^l is assigned, the continuation for the object $o_{t-1}^{k_2}$ is searched on the original frame in the place limited by a bounding box of the object o_t^l . Then the found region is cut off from o_t^l and assigned to $o_{t-1}^{k_2}$.

3.4 Pattern Recognition

The input: trajectories gathered from the tracking algorithm. *The output*: a recognized vehicle motion class. The trajectories returned by the tracking algorithm consist of a sequence of points which correspond to the positions of object centroids $o_{t_{begin}}^k \ldots o_{t_{end}}^k$ representing k-th moving vehicle. Because we wanted to make the representation independent of the position we propose to use eq. 3. to represent trajectories.

$$T_i = \left[\arctan \frac{y_{k,i+1} - y_{k,i}}{x_{k,i+1} - x_{k,i}} \right] \tag{3}$$

where $x_{k,i}$ and $y_{k,i}$ correspond to coordinates of the centroid of the object o_i^k on the frame i. A sequence of T_i is given as input of the pattern recognition algorithm.

To model vehicle behaviours on crossroads the HMM – Hidden Markov Model has been applied. In order to recognize four defined motion patterns four prelearned HMMs for each recognized motion class were used. They were learned using Baum-Welch algorithm. Hidden states of HMM are not directly defined and their number is appointed during training HMMs for each behaviour pattern. For each hidden state the probability of transition between neighbouring states and the probability of generating each observations is different.

Pattern recognition starts with the input sequence T acquisition. Next, for each HMM the probability of the sequence T generation by each HMM is calculated using forward-backward algorithm. After training HMMs the system is ready for recognition. As the output of the system class label is returned based on the HMM with the maximum probability.

4 Experimental Studies

The aim of performed experiments was to evaluate the efficiency of the system. All video sequences show an isometric view on crossroads and were recorded in different traffic intensity and different weather conditions. They were 2-17 minutes long, recorded by a low-quality webcam with low resolution. They are described in Table 1. The frames characterize mutual obscure of cars and low size of each vehicle which caused a lot of difficulties for the tracking algorithm.

Table 1. The accuracy of tracking algorithm

Name	duration	traffic intensity	*Acc*
ws_20111023_d	9min 10s	low	0,5766
ws_20111010_d	3min 41s	medium	0,5151
ws_20111005_d	2min 26s	huge	0,3880
ws_20111007_d	5min 24s	huge	0,4281
ws_20111017_d	6min 3s	huge	0,4481

Experiment1 – Evaluation of the Tracking Algorithm. The accuracy of our tracking algorithm was tested using clips presented in Table 1. Every clip was separately processed. It allows to obtain a set of trajectories recorded on each video sequence. The results of the tracking algorithm were analyzed frame by frame and if the current trajectory referred to more than one vehicle at the same time, it was abandoned. In this way a set of testing trajectories was obtained for the next tests. The accuracy *Acc* of the tracking algorithm is defined as the ratio of correctly recognized trajectories to all detected trajectories on each video sequence (eq. 4):

$$Acc = \frac{N_0}{N} \tag{4}$$

Table 2. The results of motion pattern recognition

Name	Class	*Precision*	*Recall*	$F - Score$
ws_20111023_d	straight	0.9716	0.8879	1.8557
	turn left	0.8245	0.9591	1.7734
	turn right	0.8333	0.8695	1.7020
	turn back	0.7500	1.0000	1.7143
ws_20111010_d	straight	0.9139	0.7727	1.6748
	turn left	0.5116	0.7586	1.2222
	turn right	0.6875	0.8461	1.5172
	turn back	0.5714	0.8000	1.3333
ws_20111005_d	straight	0.9411	0.8000	0.8648
	turn left	0.3333	0.4000	0.7272
	turn right	0.1818	0.4000	0.5000
	turn back	–	–	–
ws_20111007_d	straight	0.7916	0.7755	1.5669
	turn left	0.5333	0.6153	1.1427
	turn right	0.3333	0.2105	0.5161
	turn back	0.7000	0.8750	1.5556
ws_20111017_d	straight	0.9263	0.9151	1.8413
	turn left	0.6774	0.8076	1.4736
	turn right	0.5000	0.3000	0.7500
	turn back	–	–	–

where: N_0 is the number of trajectories correctly detected on video sequence, N is the number of all trajectories recognized for the sequence. The results of this test are shown in the Table 1. The tracking algorithm was optimized to track many vehicles at the same time and deal with some possible scenarios such as partial obscure of foreground objects by a static object like lantern or splitting and merging two moving vehicles. But there are situations where the proposed algorithm fails, for example when many vehicles stay very close to each others, waiting for the change of traffic lights. This is the reason why the accuracy is not very high.

Experiment2 – Evaluation of the Pattern Recognition Method. In order to calculate the accuracy of pattern recognition one HMM for one motion pattern class was prepared using trajectories gathered from the sequence ws_20111023_d as the training set. Next, the classes recognized by our recognition algorithm were compared with the manually annotated sequences. In this test only the correct trajectories were taken (not rejected in the previous stage). Based on this results we computed three measures: *Precision*, *Recall* and $F - Score$. The results of this experiment are shown in Table 2. The highest value of $Fscore$ was obtained for the sequence ws_20111023_d, which was used in the training process. The weakest recognized class is the class – *turn back* but this class of motion pattern was the most rarely represented in the sequences. For two video sequences – (ws_20111017_d and ws_20111005_d) there was no pattern from this class. Very small number of trajectories was registered for the class *turn left* and *turn right* in the video sequence `ws_20111005_d`. It was reflected in the small value of *Precision* and *Recall* for these classes. The best recognised class was *straight*. This is the result of the highest number of learning examples in the video sequence.

5 Conclusions

Although at the current state the results are far of expected by us, we think that the system has a potential for further improvement. We have to take in mind that the system acts without any semantics about the scene. Also a perspective view in video sequences aggravated trajectories acquisition. The experimental study has shown that the accuracy of the whole system is mainly dependent on the vision module responsible for tracking of objects and trajectories acquisition. Currently the system is based on trajectories for particular objects (blobs). It does not recognize a car. In the future development of the system we plan to search particular vehicles by implementing a new descriptor. Merging of vehicles in the frame causes many problems. The solution could be a recognition of the vehicle based on a vehicle top separated by a rear window.

References

1. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE Trans. Pattern Anal. Mach. Intell. 28(9), 1450–1464 (2006)
2. Masoud, O., Papanikolopoulos, N.P., Member, S.: A novel method for tracking and counting pedestrians in real-time using a single camera. IEEE Transactions on Vehicular Technology 50, 1267–1278 (2001)
3. Shi, X., Ling, H., Blasch, E., Hu, W.: Context-driven moving vehicle detection in wide area motion imagery. In: 21st International Conference on Pattern Recognition (ICPR), pp. 2512–2515 (2012)
4. Lethaus, F., Baumann, M.R.K., Köster, F., Lemmer, K.: Using pattern recognition to predict driver intent. In: Dobnikar, A., Lotrič, U., Šter, B. (eds.) ICANNGA 2011, Part I. LNCS, vol. 6593, pp. 140–149. Springer, Heidelberg (2011)
5. Hervieu, A., Bouthemy, P., Le Cadre, J.P.: A hmm-based method for recognizing dynamic video contents from trajectories. In: Image Processing, ICIP (2007)
6. Dillencourt, M.B., Samet, H., Tamminen, M.: A general approach to connected-component labeling for arbitrary image representations. J. ACM 39(2), 253–280 (1992)
7. Penoyer, R.: The alpha-beta filter. C Users Journal (1993)

Automatic Route Shortening Based on Link Quality Classification in Ad Hoc Networks

Zilu Liang and Yasushi Wakahara

Graduate School of Engineering, The University of Tokyo,
Yayoi 2-11-16, Bunkyo-Ku, Tokyo, Japan, 113-8658
z.liang@cnl.t.u-tokyo.ac.jp, wakahara@nc.u-tokyo.ac.jp

Abstract. The highly dynamic topology of an ad hoc network often causes route redundancy in the network. Several route shortening methods have been proposed to eliminate the redundancy. However, the existing works do not consider intensively the quality of the shortening links; unstable shortening links may degrade the network performance by causing unnecessary control overhead. In this paper, we seek to enhance the performance of route shortening through intensive consideration of the quality of shortening links. In our proposed SVM-ARS, all potential shortening opportunities are classified into preferred shortenings and non-preferred shortenings, and only preferred shortenings are executed in practice. The classification is achieved by Support Vector Machine (SVM). We compared SVM-ARS with a node mobility prediction model UMM and the Geographic Automatic Route Shortening (GARS) protocol. The simulations results confirm that our proposal significantly outperforms UMM and GARS through reducing the control overhead.

Keywords: Ad hoc network, routing, machine learning, support vector machine.

1 Introduction

A wireless ad hoc network is a self-organized network without the aid of any fixed infrastructure. In an ad hoc network, there can be some route redundancy due to the node mobility. In extreme cases, the redundant route may even form a loop and thus significantly degrade the performance of the network. Fig. 1 shows an example of the route redundancy where node C moves into the transmission range of node A and node B thus becomes a redundant node on the route {S,A,B,C,E,D}.

Automatic route shortening is considered a promising way to remove the redundancy in the network and to further improve the scalability of routing protocols. In existing algorithms, route shortening is performed as long as route redundancy is detected. However, if the shortening link is not stable, the execution of the shortening would lead to link failure and trigger a route recovery afterwards. We believe that such unstable shortenings is not worthy of execution in practice. The problem of interest is how to judge the quality of shortening

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_33, © Springer International Publishing Switzerland 2014

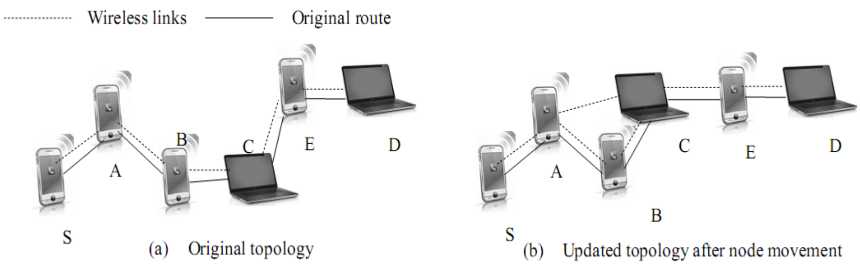


Fig. 1. An example of route redundancy due to node mobility

links and classify the potential shortenings into preferred ones and non-preferred ones. In this paper, we seek to realize the link quality prediction and classification using support vector machines (SVMs) [16]. The rationale of using SVM will be discussed later.

The rest of the paper is organized as follows. In the next section we discuss related works. In Section 3, we formulate the problem that has been addressed. In Section 4, we present the design of the route shortening algorithm SVM-ARS. In Section 5, the details on the off-line training for SVM are explained. Evaluation based on simulation results is described in Section 6. In the last section we draw conclusions.

2 Related Works

Although the ad hoc network routing problem has triggered wide research interests, the research work on the route shortening in ad hoc networks is somewhat limited. Existing methods can be divided into power-detection-based [11] and overhearing-based [1, 2, 4–6]. These methods either introduce extra control overhead [2, 4] or leads to false dection [1, 11], or simply fail to ensure satisfying data delivery ratio [5]. Xu et al. proposed a dynamic model [10] that can be used to analyze or evaluate route shortening algorithms. However, the strong assumptions on node distribution and mobility model greatly limit the application of this model.

In this paper, we present an automatic route shortening method called SVM-ARS based on support vector machine (SVM). Using SVM, nodes can predict the future performance of potential route shortenings based on their past experiences, and then decide whether to execute it or not. In this way, SVM-ARS outperforms the existing methods.

3 Problem Formulation

The core problem that we need to address is how to classify a shortening link to either preferred or non-preferred one by predicting its quality. Although several link quality prediction methods have been proposed based on estimating link life time [21, 22], these proposals are either lacking adaptability to varied scenarios,or too complex and computationally-intensive.The rationale of applying a machine learning approach lies in the following three aspects: firstly, to avoid the judgment of link quality based on the explicit mobility model of nodes; secondly, to eliminate the dependency on the properties of lower layers, such as the radio wave propagation model and the directionality of antenna used; thirdly, the performance of the machine learning based method can be further improved through refining training procedure and calibrating the parameters. Among various machine learning mechanisms, we selected the linear binary classifier SVM. On the one hand, SVM smartly converts non-linear classification into linear one by mapping vectors into a high-dimension feature space, and then solves the non-linear problem in a linear way. Due to this characteristic, SVM outperforms a lot other machine learning methods in terms of computational cost. On the other hand, SVMs are designed to minimize the structural risk by minimizing an upper bound of the generalization error rather than the training error [16]. In this way, they can provide better generalization capabilities. Therefore SVMs are particularly suitable for our problem. In the following sessions, we formulate the link classification problem into a machine learning problem using SVM.

3.1 Input Features and Output Label

The link quality related to node mobility is mainly characterized by the distance between two nodes [17] and the velocity of nodes [15], etc. Suppose {N$_1$,..., N$_{i-1}$, N$_i$,N$_{i+1}$,..., N$_{\max}$} is the concerned route in the network where N$_{i-1}$, N$_i$, N$_{i+1}$,... are successive nodes on the route, the following parameters are selected as input features:

- $|N_{\mathrm{i}-1}N_{\mathrm{i}}|$, $|N_{\mathrm{i}-1}N_{\mathrm{i}+1}|$, $|N_{\mathrm{i}}N_{\mathrm{i}+1}|$, where $|XY|$ denotes the distance between nodes X and Y.
- $v_{\mathrm{i}-1}$, v_{i}, $v_{\mathrm{i}+1}$ which are the speeds of nodes N$_{i-1}$, N$_i$, and N$_{i+1}$ respectively.
- $\theta_{\mathrm{i}-1}$, θ_{i}, $\theta_{\mathrm{i}+1}$ which are the moving directions of nodes N$_{i-1}$, N$_i$, and N$_{i+1}$ respectively.

The output label y is either -1 (preferred shortening) or 1 (non-preferred shortening). A shortening link is characterized by the input feature vector and the corresponding output label represented by Equation. 1.

$$\overrightarrow{x} = \big\{|N_{i-1}N_i|, |N_{i-1}N_{i+1}|, |N_iN_{i+1}|, v_{i-1}, v_i, v_{i+1}, \theta_{i-1}, \theta_i, \theta_{i+1}, \big\}, y \in \big\{-1, 1\big\} \tag{1}$$

3.2 Decision Function

Given a set of training data $\{(\overrightarrow{x_1}, y_1)$,...,$(\overrightarrow{x_n}, y_n)\}$, where $\overrightarrow{x_i} \in \Re^M$, M is the dimension of the feature vector, and $y \in \big\{-1, 1\big\}$, SVMs seek to construct an optimal classification function, a separating hyperplane in the feature space, to classify the training data. In our proposal, the dimension of the feature space is $M = 9$. The constructed classification function will be used as the decision function to decide whether a route shortening should be executed in practice. The optimal classification function can be obtained by solving the following quadratic optimization problem:

$$\min_{\overrightarrow{w},b} \frac{1}{2} < \overrightarrow{w}, \overrightarrow{w} > +C\sum_{i=1}^{n} \xi_i \tag{2}$$

subject to the constraints:

$$y_i\big(< \overrightarrow{w}, \overrightarrow{x_i} > +b\big) \geq \big(1 - \xi_i\big), \forall i = 1, ..., n \tag{3}$$

where $\overrightarrow{w}$ is the normal vector to the hyperplane, $\xi_i \geq 0$ for $i = 1, ..., n$ are slack variables introduced to handle the non-separable case [16], the constant $C > 0$ is a parameter that controls the trade-off between the separation margin and the number of training errors, and b is the offset of the hyperplane from the origin along the normal vector $\overrightarrow{w}$. Using the Lagrange multiplier method, one can obtain the following Wolfe dual form of the primal quadratic programming problem:

$$\min_{\alpha_i, i=1,...,n} \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j < \overrightarrow{x_i}, \overrightarrow{x_j} > -\sum_{i=1}^{n} \alpha_i \tag{4}$$

subject to the constraints:

$$0 \leq \alpha_i \leq C, \forall i = 1, ..., n, \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5}$$

where α_i for $i = 1, ..., n$ are the Lagrange multipliers. Applying the kernel tricks and solving the dual problem , the decision function can be represented by the following linear form in the feature space :

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i K(\overrightarrow{x}, \overrightarrow{x_i}) + b) \tag{6}$$

4 The Design of SVM-ARS

Before describing the proposal in detail, we present several assumptions here. We assume that all nodes in the network have the location information of themselves, which can be obtained through e.g. GPS, wireless signal measures [12], range measurements based on infrared [18] or location estimation based on connectivity information (e.g. hop count [13]). Moreover, the position information can be updated adaptively to node mobility rather than regularly [15]. All nodes also have the geographical information of their neighbours. The details on how to

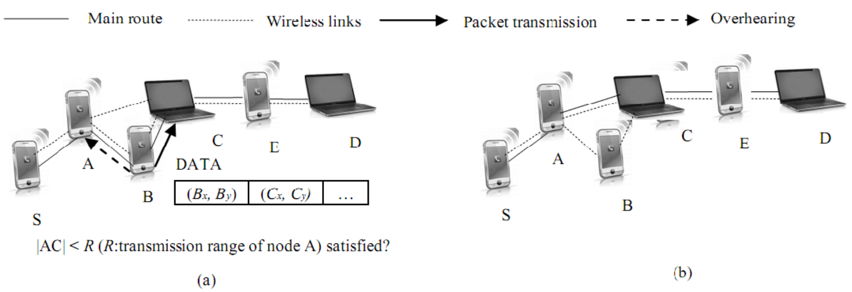


Fig. 2. Principle of SVM-ARS

obtain and manage the position information of neighbour nodes is beyond the scope of this paper. Related works with these details can be found in [14, 15]. All nodes in the network are required to operate in a promiscuous mode. Standard IEEE 802.11 [3] is used as the MAC layer protocol.

4.1 Detection of Route Redundancy

The first step of our proposed SVM-ARS is similar to the geographic route shortening method GARS [6]. When a node sends a packet, it is required to attach its own geographic information as well as the location of the next hop into the packet header in GARS. Suppose {S, A, B, C, D} is the concerned route in the scenario shown in Fig. 2. The location information of B and C, denoted by (B_x, B_y) and (C_x, C_y) respectively, is attached into the packet header when B sends the packet to C.

Since all the nodes in the network are operating in the promiscuous overhearing mode, the neighbor of B, namely node A in Fig. 2(a), can overhear the transmission at B and thus acquire the location information of C which is two-hop downstream of A. In this way, A can judge whether C has moved into its transmission range from the geography information of C. If the answer is in the affirmative, direct communication between A and C is possible and B has become a redundant node on this route, which means that a shortening process may be initiated and that the original route {S, A, B, C, E, D} can be shortened to {S, A, C, E, D}.

4.2 Classification of Shortening

In practice, all the potential shortenings can be classified into preferred ones and non-preferred ones based on certain criteria. The definition of non-preferred shortening and preferred shortening is as follows: if the shortening link breaks before any other link on this route after the execution of the route shortening, this shortening execution is considered as a non-preferred shortening; or else, it is considered as a preferred shortening. Using the decision function built through the SVM training, nodes can achieve the classification by predicting the performance of a future shortening based on their past experience. In SVM-ARS, the network system is trained offline using a pre-collected training data set. The training data set contains multiple samples; each sample is composed of input attribute vector and the corresponding label as defined in Session 3. When the route redundancy is detected in SVM-ARS, the input feature will be subtracted and input into the decision function. If the output label is -1, it means the potential shortening is preferred thus can be executed; or else the shortening is non-preferred and the execution of it may harm network performance afterwards. Take the scenario in Fig. 2(a) as an example, the route redundancy exists when $|AC| < R$. If such redundancy is detected, node A will try to judge the quality of the shortening link by inputting $\overrightarrow{x} = \{|AB|, |AC|, |BC|, v_A, v_B, v_C, \theta_A, \theta_B, \theta_C\}$ into the decision function obtained in the off-line SVM training. Providing that the output of the decision function is $y = -1$, this shortening is preferred and should be executed in the next step; or else it is non-preferred and should be ignored.

4.3 Execution of Preferred Shortening

Still using the scenario in Fig. 2 as an example, if the output of the decision function is -1 which represents a preferred shortening, node A updates its Route Table by substituting B with C on the route, and one-hop preferred shortening is achieved thereby. The original route {S, A, B, C, D} is therefore shortened to {S, A, C, D}, as is shown in Fig. 2(b)

5 Off-line Training

5.1 Training Data Collection

We implemented the SVM library LIBSVM [7] in NS-2 simulator [8] to execute the training process. The training data is collected in the following way. We run simulation in NS-2 simulator under the scenario shown in Table. 1, and all route shortenings are executed without considering the quality of shortening links. For each shortening execution, we extract the values of the 9 features representing the local environment and node mobility pattern when the shortening is executed. Note that in NS-2 a trace-file inclusing all routing information will be generated when the simulation terminates. By analysing the trace-file, we will know whether a shortening link actually has triggered a route failure in the simulation before any other links in the route. If so, the label for the features corresponding to that shortening link is 1, which represents non-preferred shortening; otherwise -1, which represents preferred shortening. We only take one sample from each run of simulation to ensure the independency among the collected samples. After running 4003 simulations, we collected a training data set with 4003 samples, among which 1994 samples are with label 1 and 2009 sample with -1.

Table 1. Simulation Scenario for Training Data Collection

Parameters	Values
Number of Nodes	200
Node Mobility Model	Random Waypoint Model
Radio Propagation Model	Two-Ray Ground Model
Network Scale	1200m Square
Maximum = Minimum Speed	1,5,10,15m/s
Pause Time	0s
Transmission Range	250m
Interference Range	250m
Number of CBR Connection	1
Packet Size	512bytes
Data Rate	64kbps
Bandwidth	2Mbps
Simulation Time	5000s

5.2 Kernel Function Selection

In order to select a proper kernel function for our problem, we study the following four widely used typical kernel functions using the collected training data: linear, polynomial, RBF, sigmoid. We use 5-fold cross-validation (CV) to avoid the over-fitting of the training data while providing good generalization. In 5-fold CV, the training set is divided into 5 subsets of equal size. In sequence, each subset is tested using the trained SVM (with a certain kernel function applied) on the other 4 subsets. We investigated the CV accuracy and the training time of each kernel function and the best performance of each kernel type is shown in Table. 2. The results indicate that RBF kernel is the best candidate for our problem. We use a "grid-search "[7] to find the values for parameters C and γ in RBF kernel which yields the best cross-validation accuracy: $C = 0.5$,$\gamma = 0.0005$. The parameters for SVM can be decided by other methods such as [9].

6 System Performance Evaluation

The proposed route shortening method SVM-ARS can be implemented into any existing routing protocol to perform route optimization. In our simulation we implemented it based on our previous work GARS [6] and evaluated the performance by ns-2 simulator. The performance of SVM-ARS is compared with the GARS, and a mobility prediction model proposed in [21] which is denoted

Table 2. Best Performance of Each Kernel Type

Kernel	Best CV Accuracy	Training Time
Linear	84.0%	22min
Polynomial	84.6%	651min
RBF	86.2%	8s
Sigmoid	50.2%	14s

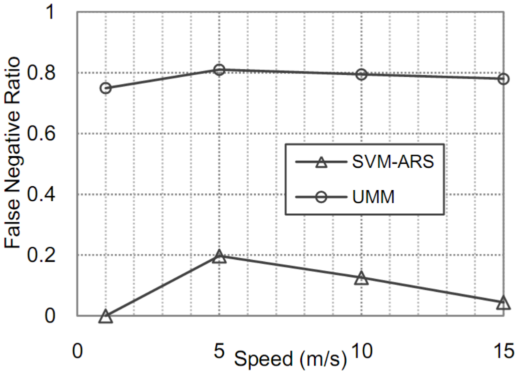


Fig. 3. False negative ratio

by UMM (UCLA Mobility Model). We first studied the classification accuracy in terms of false negative (non-preferred shortening misclassified as preferred shortening). Fig. 3 demonstrates the false negative ratio, which is the number of false negative divided by the total number of shortening executions. On average, the false negative ratio in SVM-ARS is less than 20%, whereas in UMM it is approximately 80%.

We also evaluated the performance of all these methods with respect to data packet delivery ratio and normalized control overhead. As Fig. 4(a) shows, the packet delivery ratio slightly decreases as the speed of nodes goes up for GARS and SVM-ARS, but sharply drop for UMM. SVM-ARS achieves statistically equivalent packet delivery ratio compared to GARS, and up to 32% improvement compared to UMM. In terms of normalized control overhead, SVM-ARS significantly reduces the average control overhead by 40% and 47% with respect to GARS and UMM respectively, as is depicted in Fig. 4(b).

The results verify that the proposed SVM-ARS significantly reduces the normalized control overhead compared to all other methods, and at the same time successfully ensures a satisfying data packet delivery ratio regardless of the change in the speed. The reason for the reduction of control overhead could be that the judgment on the quality of shortening link helps prevent route failure

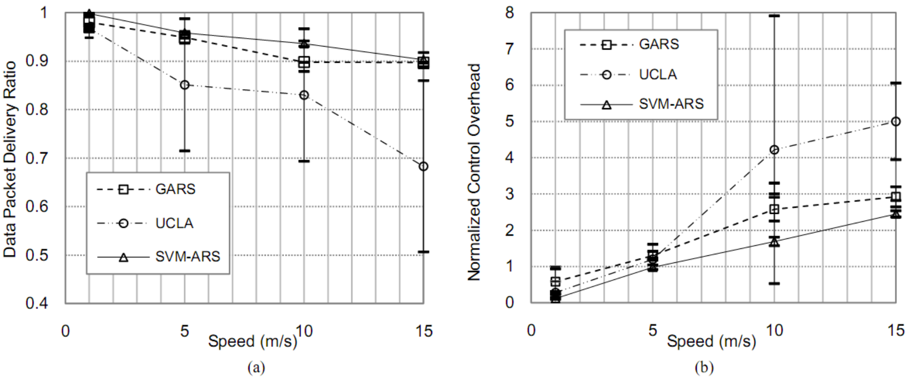


Fig. 4. Simulation results under varying speed. (a) Data packet delivery ratio; (b) Normalized control overhead.

by replacing unstable links with shorter and more stable links, and the efficiency of the judgment is verified by the low false negative ratio.

7 Conclusions

We have proposed a novel automatic route shortening method called SVM-ARS that intensively considers the quality of the shortening links. The statistic machine learning method SVM is used to predict the quality of shortening links and classify potential shortening opportunities into either non-preferred shortenings or preferred shortenings, and only preferred shortenings are executed in practice. We compared our method with GARS and a node mobility model UMM. The simulation results confirmed that SVM-ARS significantly reduces the normalized control overhead by 40% and 47% on average, while ensuring satisfying data delivery ratio. We can come to the conclusion that the resulting shorter and more stable routes in SVM-ARS are translated into an overall performance improvement of route management in an ad hoc network.

References

1. Liu, J.-S., Lin, C.-H.: RBR: Refinement-Based Route Maintenance Protocol in Wireless Ad Hoc Networks. Computer Communications 28, 908–920 (2005)
2. Cheng, R.-H., Wu, T.-K.: A highly topology adaptable ad hoc routing protocol with complementary preemptive link breaking avoidance and pat shortening mechanisms. Wireless Network 16, 1289–1311 (2010)
3. IEEE 802.11, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Std. (2007)
4. Gui, C., Mohapatra, P.: A Framework for Self-Healing and Optimizing Routing Techniques for Mobile Ad Hoc Networks. Wireless Networks 14, 29–46 (2008)

5. Liang, Z., Taenaka, Y., Ogawa, T., Wakahara, Y.: Pro-Reactive Route Recovery with Automatic Route Shortening in Wireless Ad Hoc Networks. In: Proc. of ISADS 2011 (2011)
6. Liang, Z., Taenaka, Y., Ogawa, T., Wakahara, Y.: Automatic route shortening for performance enhancement in wireless ad hoc networks. In: The 13th Network Software Conference (2011)
7. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for supportive vector machines. ACM Trans. on Intel. Sys. and Tech. 2, 1–27 (2011)
8. NS Notes and Document, http://www.isi.edu/nsnam/ns/
9. Chapelle, O., Bousquet, O., Mukherjee, S.: Choosing Multiple Parameters for Support Vector Machines. Machine Learning 46, 131–159 (2002)
10. Xu, J., Li, Q.-M., Zhang, H., Liu, F.-Y.: Model and Analysis of Path Compression for Mobile Ad Hoc Networks. Computers and Electrical Engineering 36, 442–454 (2010)
11. Chen, C.-W., Wang, C.-C.: A Power Efficiency Routing and Maintenance Protocol in Wireless Multi-Hop Networks. The Journal of Systems and Software 85, 62–76 (2012)
12. Nguyen, X., Jordan, M., Sinopoli, B.: A kernel-based learning approach to ad hoc sensor network localization. ACM Trans. Sensor Networks 1, 134–152 (2005)
13. Feng, V.-S., Chang, S.Y.: Determination of wireless networks parameters through parallel hierarchical support vector machines. IEEE Trans. Parallel and Distributed Systems 23 (2012)
14. Fiore, M., Casetti, C.E., Chiasserini, C.-F., Papadimitratos, P.: Discovery and verification of neighbour positions in mobile ad hoc networks. IEEE Trans. Mobile Computing 12 (2013)
15. Chen, Q., Kanhere, S.S., Hassan, M.: Adaptive position update for geographic routing in mobile ad hoc netwoks. IEEE Trans. Mobile Computing 12 (2013)
16. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer (2000)
17. Akyildiz, I.: Wireless sensor networks: a survey. Computer Networks 38, 393–422 (2002)
18. Roberts, J., Stirling, T., Zufferey, J., Floreano, D.: 2.5D infrared range and bearing system for collective robotics. In: Proc. of IEEE/RSJ IROS, pp. 3659–3664 (2009)
19. Chen, Y.-W., Lin, C.-J.: Combing SVMs with various feature selection strategies. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.) Feature Extraction. STUDFUZZ, vol. 207, pp. 315–324. Springer, Heidelberg (2006)
20. Fan, R.-E., Chen, P.-H., Lin, C.-J.: Working set selection using second order information for training SVM. Journal of Machine Learning Research 6, 1889–1918 (2005)
21. Su, W., Lee, S.-J., Gerla, M.: Mobility prediction and routing in ad hoc wireless networks. International Journal of Network Management 11, 3–30 (2001)
22. Guo, Z., Malakooti, S., Sheikh, S.: Multi-objective OLSR for proactive routing in MANET with delay, energy, and link lifetime predictions. Applied Mathematical Modelling 35, 1413–1426 (2011)

A Real-Time Approach for Detecting Malicious Executables

Samir Sayed[1,3,4], Rania R. Darwish[2], and Sameh A. Salem[1]

[1] Department of Electronics, Communications, and Computers Engineering,
Helwan University, Cairo, Egypt
[2] Department of Mechanical Engineering, Mechatronics
Helwan University, Cairo, Egypt
[3] Department of Electronic and Electrical Engineering,
University College London, London, UK
[4] Egyptian Computer Emergency Response Team (EG-CERT),
National Telecom Regulatory Authority (NTRA), Cairo, Egypt
{samir_abdelgawad,sameh_salem}@h-eng.helwan.edu.eg

Abstract. In this paper, we develop a real-time algorithm to detect malicious portable executable (PE) files. The proposed algorithm consists of feature extraction, vector quantization, and a classifier named Attribute-Biased Classifier (ABC). We have collected a large data set of malicious PE files from the Honeynet project in the EG-CERT and VirusSign to train and test the proposed system. We first apply a feature extraction algorithm to remove redundant features. Then the most effective features are mapped into two vector quantizers. Finally, the output of the two quantizers are given to the proposed ABC classifier to identify a PE file. The results show that our algorithm is able to detect malicious PE file with 99.3% detection rate, 97% accuracy, 0.998 AUC, and less than 1% false positive rate. In addition, our algorithm consumes a fraction of seconds to test a portable executable file.

Keywords: Portable executables, malicious detection, data mining, vector quantization.

1 Introduction

Cyber criminals threaten critical information infrastructures, and they negatively affect the national economy. For example, 24,000 files of weapons under development for the US Department of Defense (DoD) are stolen from a US defense contractor [1]. the DoD has dedicated at least 500 million USD for cyber security research. The UK government recently committed £650 million in addressing the growing cyber risks [2]. In 1990's, malware programs were manually analyzed because of their limited forms and numbers [3]. However, with the enormous increase in the number of internet users, hackers attack user's privacy through ever increasing malware (malicious software) such as Botnets, Rootkits, Worms, Trojan horses, Viruses, Spams, Adwares, and Social engineering on social networks [4]. It is impossible to depend on manual analysis to cope with the development of the new malware files. The most existing techniques such as intrusion detection systems (IDS) and antiviruses are based on signature-based algorithms to recognize malware programs [5, 6]. However, metamorphic and polymorphic

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_34, © Springer International Publishing Switzerland 2014

techniques are commonly used to bypass these algorithms [7]. Therefore, automatic real-time techniques have become urgent to detect malware files. Meanwhile, Portable Executable (PE) files play a crucial role in all of the Microsoft's operating systems and have become manifest in most malware files. Hence, this work is motivated towards developing real-time algorithm for malicious PE files detection. The proposed analysis is resting on malware detection through exploiting data mining algorithms.

There are limited research on automatically malicious executable detection using data mining and machine learning techniques [8–13]. Schultz *et al.* [8] presented a framework using deferent data mining techniques to detect malicious executables. The authors applied three algorithms: a learner based on inductive rule, a probabilistic predictor, and a multi-classifier. The proposed system outperforms signature-based systems with accuracy of 96.88%. Their system focused mostly on viruses and trojans. Wang *et al.* [9] proposed virus detection by using Decision Trees classifier. The authors used a vector of OpCode byte and the first byte of the operand as features. The results show that the performance of the proposed system by using the first OpCode byte only is better than using both the OpCode byte and the first byte of the operand. The work in [8] is improved by Kolter *et al.* [10]. The authors use the best 500 n-grams and construct several classifiers including Instance Based Learner, Support Vector Machines (SVM), Naive Bayes, Decision Trees, and Boosting to identify to PE files. They results show that the best performance is obtained by boosted decision tree J48 with an area under the ROC curve of 0.95. Perdisci *et al.* [11] extend the work in [8,10] taking into account packed PE files. They developed system called Malware Collection Booster (McBoost). It consists of two classifiers (C1 and C2) where C1 is used for packed files and C2 is used for unpacked files. They developed universal unpacker to obtain the hidden from the PE file. The results show that McBoost has an accuracy of 87.3% and AUC equal to 0.977. However, the performance of the McBoost will be degraded if the proposed unpacker fails to extract the hidden codes in the PE file. Ye *et al.* in [12,13] presented two algorithms based on objective-oriented association (OOA) mining to identify malware. In [12], they obtained a detection rate of 92%. They proposed another system called Malware Detection System (IMDS) [13] which consists of three components: PE parser, OOA rule generator, and classifier based on generated rules. They show that the proposed system outperforms well known anti-viruses with detection rate of 93.8%. Their system is incorporated into KingSoft's Anti-Virus software.

Different from discussed studies, our work is based on designing a real-time algorithm to detect whether a PE file is benign or malicious. The proposed algorithm consists of feature extraction, vector quantization, and a classifier named Attribute-Biased Classifier (ABC). We have collected a large data set of malicious PE files from the Honeynet project in the EG-CERT [14] and VirusSign [15] to train and test the proposed system. We first apply a feature extraction algorithm to remove redundant features. Then the most effective features are mapped into two vector quantizers. Finally, the output of the two quantizers are given to the proposed ABC classifier to identify a PE file. The experimental results show that our algorithm is able to detect malicious PE file with 99.3% detection rate, 97% accuracy, 0.998 Area Under the ROC curve (AUC) [16], less than 1% false positive rate. In addition, our algorithm takes a fraction of seconds (0.0011 sec.) to scan PE file. Therefore, the proposed system is suitable for real-time systems.

The rest of this paper is organized as follows. Section 2 gives the system model, feature extraction algorithm, and vector quantization. The proposed Attribute-Biased Classifier (ABC) is presented in 3. Simulation results of the proposed system are presented and discussed in Section 4, followed by conclusions and future work in Section 5.

2 System Model

Referring to Figure 1, the proposed system comprises three main components: feature selection, vector quantization, and classification stage. The input PE files first pass to PE parser. The parser is a set of custom python scripts used to construct the attributes or features from the PE header [17]. These attributes are passed with file entropy to the feature select algorithm to select the best features. It should be noted that the features extracted are static properties of the PE files and there is no need to execute the PE files.

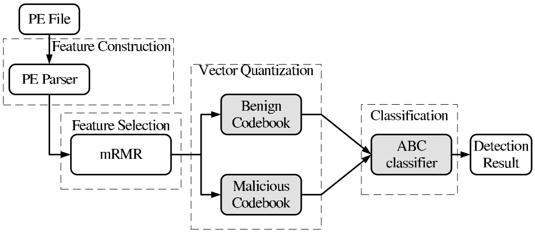


Fig. 1. Portable Executable detection system

In this research 19 attributes are constructed from header and optional header with file entropy of PE files as shown in table 1. These attributes are used to identify whether a PE file is benign or malicious.

Table 1. Table of features extracted from PE header

	Feature Name	Type
FILEHEADER	TimeDateStamp	date-time
	NumberOfSections & PtrToSymTable & CHAR-RELOCS-STRIPPED	integer
	CHAR-BYTES-RESERVED-LO & CHAR-BYTES-RESERVED-HI	integer
OPTIONALHEADER	LV-MAJ-MIN & OS-MAJ-MIN & IMAGE-MAJ-MIN	integer
	SizeOfCode/SampleSize & SizeOfInitializedData/SampleSize	float
	SizeOfUninitializedData/SampleSize & SizeOfImage/SampleSize	float
	SizeOfHeaders/SampleSize & AddressOfEntryPoint/SampleSize	float
	BaseOfCode/SampleSize & BaseOfData/SampleSize & NumberOfRvaAndSizes	float

2.1 Feature Selection

Feature extraction is a preprocessing step for many pattern recognition and machine learning systems. This step usually encompasses feature construction and feature selection. Feature construction is achieved by PE parser as shown in Figure 1. The output of the PE parser is 19 features given in table 1. Features selection is a process of selecting the most effective subset of features to minimize the training time and to improve the detection ratio and accuracy of the classifier. In this paper, min-Redundancy Max-Relevance (mRMR) feature selection algorithm [18] is used to select the best subset of features. The main idea behind the mRMR algorithm is that relevant and redundant features are considered simultaneously. The min-Redundancy is obtained by selecting the features which are maximally dissimilar to each other and ignoring the features which are very correlated. The redundancy is measured by the average of all mutual information values $I(f_i; f_j)$ between feature f_i and feature f_j in the set $\mathcal{S}$ and is calculated as follows:$R(\mathcal{S}) = \frac{1}{|S|^2} \sum_{f_i, f_j \in \mathcal{S}} I(f_i; f_j)$. The Max-Relevance is achieved by selecting features which are highly relevant to the target class. In this paper, we have two classes which are benign class and malicious class. The relevance of feature f_i in a set $\mathcal{S}$ with respect to a class c is calculated as follows:$D(\mathcal{S}, c) = \frac{1}{|S|} \sum_{f_i \in \mathcal{S}} I(f_i; c)$. where $I(f_i; c)$ is the mutual information between feature f_i and the class c. The mRMR criterion is combining feature redundancy $R(\mathcal{S})$ and feature relevance $D(\mathcal{S}, c)$ given above to obtain a good subset of features $\mathcal{S}^*$. It is given as follows:$\mathcal{S}^* = \arg\max_{\mathcal{S}}(D(\mathcal{S}, c) - R(\mathcal{S}))$

2.2 Vector Quantization

Vector Quantization (VQ) [19, 20] is one of well known techniques commonly used in image compression and speech recognition. VQ is based on three main steps: codebook design, encoding, and decoding. In this paper Linde-Buzo-Gray (LBG) algorithm [21] is used for designing two different codebooks: benign and malicious codebooks. The LBG algorithm is one of the most cited VQ algorithms used to generate VQ codebook. The LBG algorithm is explained later in this section. In the training phase, we form training matrices of the best features selected by the mRMR algorithm. These matrices are then applied to the LBG algorithm to generate the best codebook describing the matrices used in the classification phase. Let $\mathcal{F} = \{F_1, F_2, ..., F_\ell, ..., F_L\}$ be the training matrix, and $F_\ell = \{f_{\ell,1}, f_{\ell,2}, ...f_{\ell,N}\}$ is a vector of selected features from the ℓ^{th} PE file. L is the number of PE files used in the training phase to design the codebooks of both benign and malicious vector quantizers. N is the vector length of selected features. Codebook design is performed by LBG algorithm resulting in codebook $\mathcal{C}$ with size $M \times N$, where M is the length of each vector and N is the number of codevectors. In this paper, we assume that the benign and malicious codebooks have the same size $M \times N$. A feature $f_{\ell,n}$ from the vector of selected features F_ℓ is approximated by a codevector $C_n = \{c_{1,n}, c_{2,n}, ..., c_{m,n}, ..., c_{M,n}\},\ n \in N,\ m \in M$ if the quantization distortion $D_{q,n} = \frac{1}{M} \sum_{m=1}^{M} \mid f_{\ell,n} - c_{m,n} \mid^2$ is minimum. Therefore, the design problem is to construct the appropriate codebooks that best represent both benign and malicious training sets given $\mathcal{F}$, M, and N such that $D_{q,n}$ of the n^{th} feature is minimum. The LBG algorithm used to solve this minimization problem is given as follows:

Step 1: *Given $\mathcal{F}$, small number ϵ, and $m = 1$, we obtain the initial values*

$$c^*_{1,n} = \frac{1}{L}\sum_{\ell=1}^{L} f_{\ell,n} \quad D^*_{q,n} = \frac{1}{L}\sum_{\ell=1}^{L} \mid f_{\ell,n} - c^*_{1,n} \mid^2, \quad n \in N$$

Step 2: *Split the codevectors for $i = 1, 2, ..., m$ as follows: $c^{(0)}_{i,n} = (1+\epsilon)c^*_{m,n}$, $c^{(0)}_{i+m,n} = (1-\epsilon)c^*_{m,n}$, and set $m = 2m$*

Step 3: *Set iteration index $j = 0$, and $D^{(0)}_{q,n} = D^*_{q,n}$, then*

1. *Get the minimum value of $\mid f_{\ell,n} - c^{(j)}_{i,n} \mid$ for $\ell = 1, 2, ...L$, $i = 1, 2, ...m$, $n = 1, 2, ..., N$. Set $Q(f_{\ell,n}) = c^{(j)}_{i^*,n}$, where $i^* \in m$ is the index achieving the minimum.*
2. *Update the codevector $c^{(j+1)}_{i,n}$:*

$$c^{(j+1)}_{i,n} = \frac{\sum_{Q(f_{\ell,n})=c^{(j)}_{i,n}} f_{\ell,n}}{\sum_{Q(f_{\ell,n})=c^{(j)}_{i,n}} 1} \quad \textit{and set } j = j+1$$

3. *Calculate $D^{(j)}_{q,n} = \frac{1}{L}\sum_{\ell=1}^{L} \mid f_{\ell,n} - Q(f_{\ell,n}) \mid^2$*
4. *Go to (1) if $\frac{D^{(j-1)}_{q,n} - D^{(j)}_{q,n}}{D^{(j)}_{q,n}} > \epsilon$*
5. *Else set: $D^*_{q,n} = D^{(j)}_{q,n}$ and $c^*_{i,n} = c^{(j)}_{i,n}$ $\forall i = 1, 2, ..., m, \quad n = 1, 2, ..., N$*

Step 4: *Repeat Steps (2) and (3) until the desired codebook $\mathcal{C}_{M\times N}$ is achieved*

3 Attribute-Biased Classifier (ABC)

This work introduces Attribute-Biased Classifier (ABC) for the purpose of PE files classification. The main idea of the proposed classifier is to identify the closeness of the attributes extracted from a testing PE file against their counterparts attributes at the designated codebooks (benign/malware codebook). Thus, the more attributes approach a certain class codebook, the more possibility of assigning the PE file to that class. When an unknown PE file is given, the ABC searches the number of attributes approaching each class codebook. When the majority of attributes from a testing vector approach the benign (malicious) codebook, the file is identified as benign (malicious) PE file. The "closeness" of a given attribute to certain class is measured by computing the dissimilarity vector against its counterpart attribute at that class codebook. Dissimilarity vector is measured by calculating the Min-Mean-Square-Error (MMSE) between each attribute and the corresponding benign and malicious codebook vectors. We then obtain benign and malicious dissimilarity vectors. The attribute which has a MMSE with one codebook is belonging to that codebook. After that the total number of attributes associated with each codebook is counted. A PE file is considered benign if the number of the testing vector attributes associated with benign codebook bigger than that with malicious codebook. Otherwise a PE is considered malicious file. If the number of attributes belongs to both benign and malicious codebooks are equal, the minimum of each dissimilarity vector is calculated. Then, a testing PE file would be assigned to a class (benign or malicious) with the smallest minimum dissimilarity value. The pseudocode of the proposed ABC algorithm is given in algorithm 1. For a testing PE file, let $\mathcal{V}_t = \{v_1, v_2, ..., v_N\}$ be the vector of the most effective N attributes selected by the feature selection algorithm. The $\mathcal{C}_b$ and $\mathcal{C}_m$ are $M \times N$ matrices of both benign and malicious codebooks, respectively. Let D_b and D_m are the dissimilarity vectors of the test vector $\mathcal{V}_t$ with respect to both benign and malicious codebooks, respectively. The criteria used to measure the dissimilarity is the MMSE as mentioned above.

Algorithm 1. Attribute-Biased Classifier (ABC) algorithm

```
$AC_b \leftarrow 0$                                              ▷ benign counter
$AC_m \leftarrow 0$                                              ▷ malicious counter
for $i = 1 \rightarrow N$ do
    $e_b \leftarrow 0, \quad e_m \leftarrow 0$
    for $j = 1 \rightarrow M$ do
        $e_b \leftarrow e_b + \parallel \mathcal{V}_t(i) - \mathcal{C}_b(j,i) \parallel^2$ and $e_m \leftarrow e_m + \parallel \mathcal{V}_t(i) - \mathcal{C}_m(j,i) \parallel^2$
    end for
    $D_b(i) \leftarrow e_b/M$ and $D_m(i) \leftarrow e_m/M$
    if $D_b(i) < D_m(i)$ then
        $AC_b \leftarrow AC_b + 1$
    else
        $AC_m \leftarrow AC_m + 1$
    end if
end for
if $AC_b > AC_m$ then
    The PE file is benign
else if $AC_b < AC_m$ then
    The PE file is malicious
else if $AC_b = AC_m$ then
    if $\min(D_b) < \min(D_m)$ then
        The PE file is benign
    else
        The PE file is malicious
    end if
end if
```

4 Simulation Results

In this section, we present the experimental results of the proposed algorithm on a large data set of malicious and benign PE files. We gathered 2500 benign PE files from Windows machines and 3000 malicious PE files with no duplication in the data set. The malicious PE files are collected from the Honeynet project in the EG-CERT [14] and downloaded from VirusSign [15]. The data set is split into two sets, training and testing sets. The training set is used to obtain the most effective attributes from the 19 attributes given in table 1 and to design the codebooks of both malicious and benign vector quantizers. The testing set is used to measure the performance of the proposed system over unseen PE files. We have used one thousand malicious and one thousand benign PE files for training and the rest of the data set is used for testing. The procedure of selecting the training set is carried out 100 times. The results have been averaged to get a better insight into the performance of the proposed algorithm over the entire set. We performed our analysis on Intel Pentium Core 2 Duo machine, 2GHz processor, and 2 GByte RAM. Three metrics detection rate (DR), false positive rate (FPR), and overall accuracy (ACC) are used to measure the performance of our algorithm. These metrics are computed mathematically as follows:

$$DR = \frac{T_P}{T_P + F_N} \quad FPR = \frac{F_P}{F_P + T_N} \quad ACC = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

where T_P is true positive representing the number malicious PE files classified as malicious files, and true negative T_N is the number of benign PE files classified as benign files. False positive F_P stands for the number of benign files classified as malicious files, and false negative F_N is the number of malicious files classified as benign files. In addition, we obtained Area Under ROC Curve (AUC) to measure the detection accuracy of the proposed algorithm. ROC curves [16] are used in data mining to illustrate the tradeoff between true positive rate and false positive rate. As the AUC increases, the detection accuracy increases because of high rate of true positive and low rate of false positive. For example, at AUC=1, true positive rate=1, false positive rate=0, and detection accuracy=1. Figure 2 shows the relation between gain in information calculated from mRMR algorithm [18] and the 19 features/attributes constructed by the PE parser as perviously mentioned. The figure shows that the selected attributes are not similarly contributing to the classification process. Thus, the most effective five features have been selected. The selected attributes are not correlated and are more relevant to the malicious and benign classes. The design complexity of vector quantizer codebooks is also reduced, because the codebook size depends on the number of features. This also reduces the proposed classifier complexity, and as a result the processing time will reduce during testing of a PE file. The results in figure 2 is obtained from a data set including 2000 sample (1000 benign and 1000 malware).

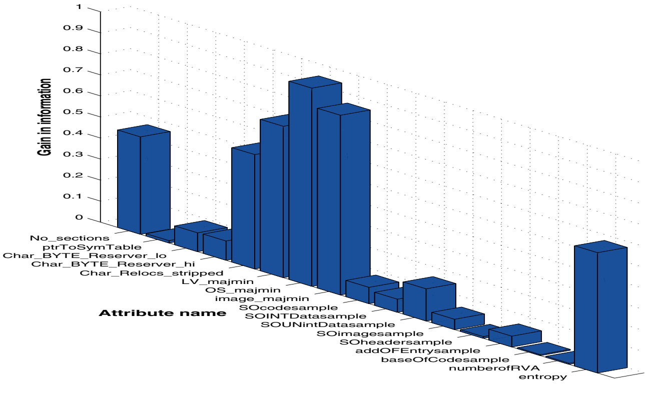
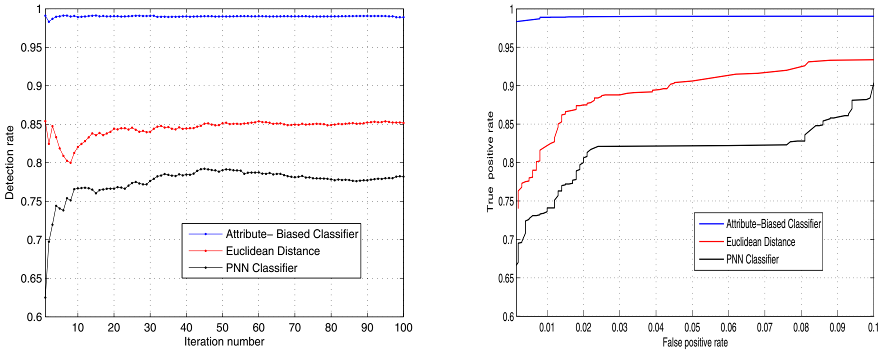


Fig. 2. Gain in information vs. attribute name

To select the optimal codebook size which achieves better performance with minimum complexity, we have obtained the relationship between the system performance in terms of detection rate and overall accuracy versus number of codebook vectors. In this research it is assumed that both benign and malicious codebooks have the same size. The optimum number of codebook vectors that yields in high detection rate and accuracy is eight vectors. The length of each vector is five coefficients as the number of selected features. Therefore, the optimum codebook size is 8×5 for both benign and malicious codebooks.

To study the performance of the proposed Attribute-Biased Classifier (ABC), we have obtained the relation between detection rate versus the number of runs as shown in figure 3(a). We compare the proposed ABC classifier with other known classifiers such as Probabilistic Neural Network (PNN) classifier [22] and Euclidean distance classifier [23]. It can be seen that, the proposed protocol has a higher detection rate (99%) than the other classifiers. Consequently, the processing time of the proposed classifier is decreased during training phase. Figure 3(b) shows the true positive rate versus false positive rate using ROC curves. The proposed ABC classifier outperforms the other classifiers. The ABC classifier has AUC more than 0.98 at false positive rate less than 2%.



(a) Detection rate vs. number of runs. (b) false positive rate versus true positive rate.

Fig. 3. Proposed approach performance

Furthermore, the effectiveness of the proposed approach veersus other recently developed techniques has been investigated. We have tabulated the AUC and the scan time. In the table we compare the proposed technique with Mc-Boost [11], Schuktz *et al.* [8] (titled Strings), Kolter *et al.* [10] (titled KM), and two well-known antiviruses: AVG [24] and Panda [25]. As shown in this table the proposed system outperforms the rest of other techniques in terms of AUC and scan time. The results in table 2 are averaged over 100 runs to effectively detect unseen malicious PE files. We can say that the proposed technique provides slight improvement in the detection accuracy compared with the other technique. However, the proposed approach has the minimum scan time of 0.0011 sec/file. Thus, this improvement in the detection accuracy becomes effective when the number of PE files to scan is several thunders. This number of PE files might be available at internet service providers (ISPs), CERTs, and other sensitive infrastructures. Therefore, the proposed system is feasible for real-time implementation because of very low processing time and high detection rate. In addition, the results obtained in table 2 show that the accuracy of the proposed protocol is better than the other protocols. The proposed algorithm requires higher processing time during the training phase than some recently developed techniques. This is because of the codebook design.

Table 2. Proposed algorithm vs. different techniques.

Technique	Classifier	AUC	DR%	Accuracy%	Scan-time (sec.)
Proposed	**ABC**	**0.99**	**99.3**	**97.05**	**0.0011**
MC-Boost [11]	IBk	0.926	72.7	87.3	3.255
Strings [8]	IBk	0.927	–	96.88	5.582
KM [10]	IBk	0.977	83.3	42.9	31.973
AVG [24]	-	-	-	-	0.159
Panda [25]	-	-	-	-	0.131

5 Conclusions and Future Work

In this paper, a real-time algorithm for detecting malicious PE files is presented. As demonstrated, the proposed algorithms uses a classifier named Attribute-Biased Classifier (ABC) to identify whether the test file is benign or malicious. Experiments on large data set collected from the Honeynet project in the EG-CERT and VirusSign show that the proposed algorithm outperforms the recently developed algorithms and popular antivirus softwares, such as AVG and Panda in terms of scan time, AUC, detection rate, false positive rate, and accuracy under the testing phase. In our future work, we plan to collect more detailed information about the PE header and use it to improve the performance of the proposed protocol. We plan also to extend our work to not only detect whether a PE file is malicious or benign, but also to identify the type of malware.

References

1. Symantec Corporation: Symantec Internet Security Threat Report. Technical report, vol. 71 (2012)
2. The UK Cyber Security Strategy: Protecting and Promoting the UK in a Digital World. Technical report (2011)
3. Zhong, Y., Yamaki, H., Takakura, H.: A Malware Classification Method based on Similarity of Function Structure. In: IEEE/IPSJ 12th International Symposium on Applications and the Internet, pp. 256–261 (2012)

4. McGraw, G.M.G.: Attacking malicious code: report to the infosec research council. IEEE Softw. 17, 33–41 (2002)
5. Filiol, E.: Malware pattern scanning schemes secure against blackbox analysis. J. Comput. Virol. 2, 35–50 (2006)
6. Filiol, E., Jacob, G., Liard, M.L.: Evaluation methodology and theoretical model for antiviral behavioural detection strategies. J. Comput. Virol. 3, 27–37 (2007)
7. Song, Y., Locasto, M., Stavrou, A., Keromytis, A., Stolfo, S.: On the infeasibility of modeling polymorphic shellcode. In: Proceedings of the 14th ACM Conference on Computer and Communications Security, pp. 541–551 (2007)
8. Schultz, M., Eskin, E., Zadok, E.: Data mining methods for detection of new malicious executables. In: Proceedings of IEEE Symposium on Security and Privacy, pp. 38–49 (2001)
9. Wang, J.H., Deng, P., Fan, Y., Jaw, L., Liu, Y.: Virus detection using data mining techniques. In: Proceedings of IEEE International Conference on Data Mining (2003)
10. Kolter, J., Maloof, M.: Learning to detect malicious executables in the wild. In: Proceedings of Knowledge Discovery and Data Mining (KDD), pp. 470–478 (2004)
11. Perdisci, R., Lanzi, A., Lee, W.: McBoost: Boosting Scalability in Malware Collection and Analysis Using Statistical Classification of Executables. In: Annual Computer Security Applications Conference (ACSAC), pp. 301–310. IEEE Press, USA (2008)
12. Ye, Y., Wang, D., Li, T., Ye, D.: IMDS: Intelligent malware detection system. In: Procceedings of ACM International Conference on Knowlege Discovery and Data Mining, SIGKDD (2007)
13. Ye, Y., Wang, D., Li, T., Ye, D., Jiang, Q.: An intelligent PE-malware detection system based on association mining. Journal in Computer Virology 4, 323–334 (2008)
14. EG-CERT, `http://www.egcert.eg/cert/`
15. VirusSign, `http://freelist.virussign.com/freelist`
16. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Laboratories (2004)
17. Pietrek, M.: Peering Inside the PE: A Tour of the Win32 Portable Executable File Format (1994)
18. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3, 185–205 (2005)
19. Gray, R.M.: Vector quantization. IEEE ASSP Mag., 4–29 (1984)
20. Gersho, A., Gray, R.M.: Vector quantization and signal compression. Kluwer Academic Publishers (1991)
21. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Transactions on Communications 28, 84–95 (1980)
22. Specht, D.F.: Probabilistic Neural Networks for Classification, Mapping, or Associative Memory. In: IEEE International Conference on Neural Networks, vol. I, pp. 525–532 (1998)
23. Marcoa, V.R., Younga, D.M., Turnerb, D.W.: The Euclidean distance classifier: an alternative to the linear discriminant function. Communications in Statistics - Simulation and Computation 16, 485–505 (1987)
24. AVG Antivirus, `http://free.avg.com/`
25. Panda Antivirus, `http://www.pandasecurity.com/`

Fault Diagnosis of a Corrugator Cut-off Using Neural Network Classifier

Jerzy Kasprzyk[1] and Stanisław K. Musielak[2]

[1] Silesian University of Technology, Institute of Automatic Control, Gliwice, Poland
jerzy.kasprzyk@polsl.pl
[2] BHS Corrugated Maschinen- Und Anlagenbau GmbH, Weiherhammer, Germany
SMusielak@bhs-corrugated.de

Abstract. In this paper a proposal for solving the problem of diagnostics of cutting errors in a rotary cutoff in a corrugated board machine processing line is presented. There are many different reasons for errors, and their identification requires a sound knowledge and experience of the service staff. The authors, using their many years' experience and a huge database, have found that many sources of errors can be characterized using a probability density function (pdf). They proposed a diagnostics method based on classification of sources of disturbances using the analysis of pdf determined with a kernel density estimator. Multilayer feedforward neural network is proposed as a classifier. Classification procedure is discussed, together with research results based on data from real industrial processes.

Keywords. corrugated board, cut-off, pattern classification, neural network, kernel density estimation.

1 Introduction

Corrugated board is the product which is obtained by gluing together the alternatively placed flat and corrugated layers of paper or board. It has many advantages, such as lightness, durability, ease of packaging production and possibility of repeated processing, thus it plays a key role in the logistics related to the flow of materials and products.

The corrugated board is produced in complex production lines consisting of 12 to 15 modules and are up to 150 m long. One of important production stages is sheeting, performed in the last phase the production process, where the board sheet, 200 to 3200 mm wide, is cut into suitable formats in the cutter [1]. Cutting is a very complicated process, the accuracy of this phase of production is one of crucial elements upon which the carton's quality is based on. Therefore, it is extremely important to create adequate control and monitoring system for cutting. From the technical point of view the problem is extremely difficult, because nowadays the corrugated board production lines operate at a speed which may reach 450 m/min, and the length of cut board pieces may vary from 400 to 10000 mm. Required cutting accuracy may be about +/-

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_35, © Springer International Publishing Switzerland 2014

1 mm for formats shorter than 2000 mm. For longer formats this accuracy is assumed to amount –1 mm to 0.0005*L, where L is the length of the board piece.

In paper [2] the system for initial quality assessment of the cutting process is proposed, based on the analysis of the cutting error bivariate distributions. Such system allows the process operators for earlier detection of irregularities in rotary cutoff power transmission systems and calling an appropriate specialist service team. The aim of this paper is to describe the error diagnostics method for the cutting process, which would make it possible to find error sources automatically, and would certainly significantly simplify a service work procedure. On the grounds of many years' experience of working with the cutoffs it has been found, that cutting errors distribution allows for defining a certain number of classes, describing the causes for these errors. Hence, the error classification method was proposed, based on artificial neural network, which, on the grounds of error distribution and expert knowledge gained through many years of service work with many different cutoffs, will allow for finding with high probability the causes for improper operation of the cutoff.

The paper is organized as follows: (1) the cutting process in the rotary cutoff and cutting errors are described, attention is drawn to the need of creating an adequate system for diagnostics of these errors, (2) proposal for the solution of the problem of intelligent diagnostics of the machine is presented, (3) the problem of creating the data set for the classifier using the kernel estimators is described, (4) the issue of error causes' classification using the neural networks is discussed, (5) experimental results of the use of the proposed diagnostic method is presented. At the end the conclusions and proposals for further investigations in this field are described.

2 Rotary Cutoff

The process of cutting is performed in a rotary cutoff machine with a so-called helicut system, in which a shearing principle is obtained thanks to rotary movement of a shaft with a knife, see fig. 1. In this case, the knife is mounted spirally on the shaft. Machines of this type are employed in many industries, for example in the process of cutting a cardboard, foil, sheet metal, paper, laminate. Although the principle of operation in each case is the same, the cutting phase course, tolerance requirements, control algorithms used and technological environment where the rotary cutoff machines operate, are different.

There are two shafts in the cutoff: the bottom one and the top one, stiffly connected with gear wheels. Knife shafts are 1800 mm to 3300 mm long. The power transmission system consists of two AC asynchronous motors, 60 KW each, mounted on either side of a knife shaft. There may be a single cutoff unit (Simplex), double units (Duplex), see fig. 1, or triple units (Triplex), in one constructional frame. Achieved speeds are from 15 m/min for older units, up to 400 m/min.

The control system of the cutoff machine is a time-optimized system with a moving target point (*rendez-vous* type arrangement) [3]. This means, that the power transmission system of the cutoff should ensure reaching the required knife position according to the set velocity profile.

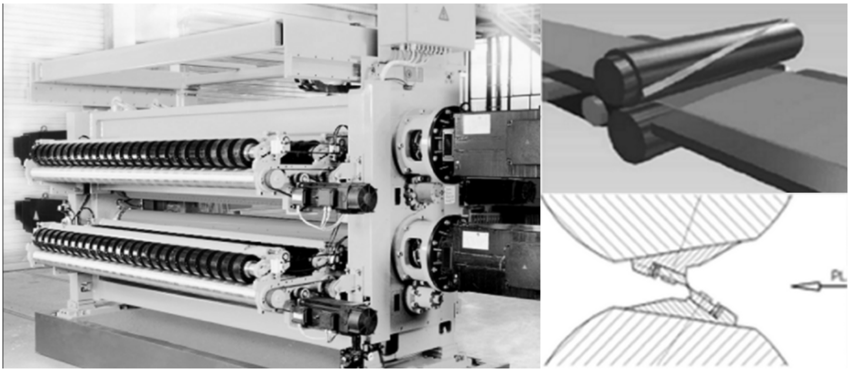


Fig. 1. Example double rotary cutoff machine (*Duplex*) and the knife shaft in a rotary cutoff

2.1 Cutting Process

Cutting phase may amount from 44 to 56 degrees of shaft revolution. In this phase there should be smooth, impact-free transition of the knife through the corrugated board, so the linear velocity of the knife must be precisely synchronized with the velocity of the corrugated board sheet. At the same time, to obtain a corrugated board of a set length, it is necessary to obtain the set position of the knife with the required velocity profile. A simplified diagram of time functions of the linear velocity of the knife for cutting the format above 1500 mm is shown in the fig. 2. Correct passage through the cutting phase requires that the knife velocity is equal to the corrugated board velocity in the production line. The exact moment and the place where the knife meets the board are precisely defined by the preset length of the corrugated board being cut.

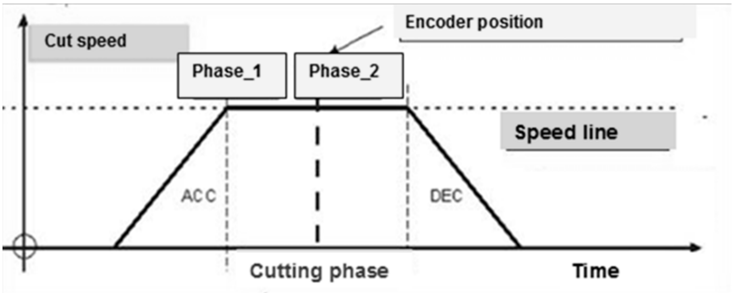


Fig. 2. Simplified diagram of knife velocity trajectory: Phase_1 - knife enters the cutting phase, Phase_2 - knife exits the cutting phase, Speed line - production line speed, ACC - knife acceleration phase DEC - knife slow down phase

There may be various causes of incorrect cutting – they may lie on the side of the cutoff and on the side of its surroundings in the production line. In the first case the errors may be caused by irregularities in the power transmission system, improper fastening of the encoder, when its zero marker does not correspond to mechanical position of the knife, improper torques of the knife shafts, improper controller settings etc. Another group of sources of errors includes for example: instabilities of the production line speed, play in the cutoff input rolls, corrugated board quality changes, etc. Fig. 3. shows some real examples of the cutting errors.

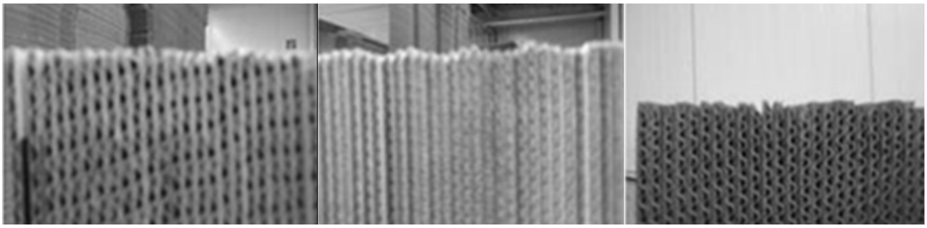


Fig. 3. Examples of the cutting errors

2.2 The Need for Intelligent Diagnostics System

A corrugator is an expensive machine, operating usually according to a three-shift scheme. The corrugator's life is usually estimated as 20-30 years of continuous operation [4]. In addition, very strict requirements for cutting accuracy are employed, about +/-1 [mm], resulting from the use of highly robotized machines used for further processing of the corrugated board. Therefore all disturbances in corrugator operation, especially cutting errors exceeding the set tolerance threshold, and standstills caused by necessary repairs, can substantially influence the economic results.

The number of possible causes of irregularities/disorders is very high. Their identification requires great knowledge and service work experience, and is often related to an arduous process of analysis of individual parts of the machine. Therefore it is very important to create an adequate tool for the cutoff error diagnostics, which with growing demands related to the cutoff cutting accuracy would allow for quick and reliable determination of causes of the cutting errors. It is especially important for older machines, because production losses caused by product quality worsening or accident conditions may be very expensive.

3 Machine Condition Diagnostics Based on Pattern Classification

Machine condition can be assessed by observing the object operation, especially its output. In the investigated case the machine condition is assessed basing on product quality observations, *i.e.*, corrugated board cutting accuracy. However, an extensive knowledge and experience is required to be able to state, on the grounds of cutting error analysis, what are the causes of resultant deviations, because, as it was mentioned, the number of possible disturbances which influence the cutting process is very high and there may be a combination of various disorders: electrical, mechanical, or resulting from the course of the production process.

Basic methods for assessment of the process or the product quality are based on the analysis of statistical parameters (this approach was proposed in [2]). Such changes, especially parameters which are out of the admissible limits, give evidence for the existence of irregularities in the process. However, they usually do not allow for finding the sources of errors, especially in such a complex process like cutting in the corrugator. That's why more advanced methods are required, like for example data mining, which allow for obtaining the information from history data very quickly, by finding the patterns [5]. Many years' specialist experience related to operation and service work with corrugated board machines and analysis of many hundreds of measurement data elements make it possible to define some classes of cutting error sources, related to the distributions of these errors. So, it was proposed to base the cutoff diagnostics on pattern classification methods. The input value for such a classifier will be the estimated probability density function (pdf) of measured cutting errors, while the output will be the class assigned to a specific group of process disturbances.

The problem of classification can be stated as follows. For given set of training data $D = \{(x_i, y_i), i = 1,\ldots,n\}$, where $x \in \mathcal{X}$ denotes a vector of pdf values calculated for the cutting errors, and $y \in \mathcal{Y}$ is a label defining the class of disturbances in the cutting process, produce a function $g : \mathcal{X} \rightarrow \mathcal{Y}$ that approximates the unknown correct mapping $y : \mathcal{X} \rightarrow \mathcal{Y}$. The goal of the learning procedure is to minimize the zero-one loss function (assigning a loss of 1 to any incorrect labeling) for a training set collected during former experiments and assumed to represent accurate examples of the mapping.

There are many classification methods, for example decision trees, nearest neighbors, neural networks, naive Bayesian classifiers and others [6][7]. Because the model of class assignment to pdf shape is strongly nonlinear and difficult to be mathematically formulated, it was decided to approach the problem basing on the use of Neural Networks (NN) in modeling the relation between pdf of the cutting errors and the class. NN classifiers [8] are commonly used in classification and decision tasks and they have been demonstrated to be a competitive alternative to traditional classifiers for many practical classification problems [9].

Classification usually comprises a set of the following operations:

- Preprocessing, *i.e.*, data preparation for the classifier;
- Feature extraction and pattern classification;
- Decision rule preparation.

In the investigated case the aim of data preparation is pdf calculation for deviations of measured pieces of board from the set value. A set of pdf values is a feature vector x. A set of training data D comprises features with disturbance class y assigned to them by an expert. But test data contain only the features. A decision rule is a function mapping feature vectors into classes.

Total number of measurement data used for learning and testing the classification method comprised over 12000 measuring series consisting of 20 to 100 measurements. Fifteen classes were distinguished in the data: class 1 corresponds to the correct cutting process, others represent different disturbances existing in the process. Classification procedure was implemented using the Matlab package with toolboxes Statistics and Neural Networks.

3.1 Corrugated Board Length Measurement

Measurement of the corrugated board sheets is performed in a classical manner, using a measuring tape. For technological reasons, no other length measuring methods are employed. In practice, this measurement is performed randomly, and the number of measurements varies from 3 to 5 for one order. Of course, this number is too small in case of a diagnostic system and does not allow for system error analysis. Therefore, when operating personnel finds that the shape of arranged stack of sheets deviates from the standard (see fig. 3), at least 30 – 50 measurements from such population should be done. Acquisition of larger number of measurement results is recommended. Especially if two modes appear in the pdf, then the number of measurements should be doubled [10]. Measurement results are influenced by line coincidence reading errors, parallax errors and experimenter errors. Maximum measurement error within the range of +/- 0.25 mm was assumed. Fig. 4 shows an example diagram of calculated deviations of board length from the set value.

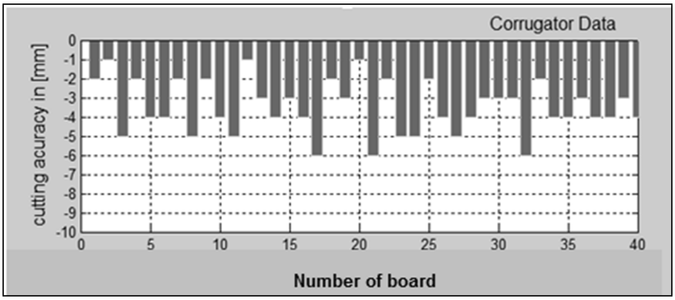


Fig. 4. Board length deviations from the set value

3.2 Pdf Estimation

Density estimation is the construction of an estimate of pdf from the observed data. In classical methods, mainly parametric approach is employed. However, in case of a corrugated board machine, because of a large number of different cutting error distributions as well as multimodality, it is difficult to find distributional assumptions a priori, so instead of parametric methods an alternative approach, non-parametric density estimation, is used.

The simplest non-parametric method is a histogram. A set, containing all samples, is divided into a number of intervals characterized by an equal width *h* and a histogram is a function that counts the number of observations falling into each of the disjoint intervals (known as bins). Histogram's advantage is simplicity and ease of interpretation in the first stage of investigation. Its main disadvantage is strong dependence of its properties on interval width *h* and sensitivity to location of the beginning of the first interval [11].

An alternative to the histogram is kernel density estimation (kde), which uses a kernel to smooth samples [12]. This will construct a smooth pdf, which in general will more accurately reflect the underlying variable.

Let $(x_1, x_2, \ldots, x_n)$ be an *iid* sample drawn from some distribution with an unknown density f. Estimating the shape of the function f can be done using kde as:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{1}$$

where $K()$ is the kernel and $h > 0$ is a smoothing parameter called the bandwidth.

It is assumed, that the kernel function is symmetrical with respect to zero $K(x) = -K(x)$ and has in this point a weak global maximum $K(0) \geq K(x)$. A range of kernel functions are commonly used: uniform, triangular, Epanechnikov, normal, and others. The choice of the kernel function K is not crucial to the accuracy of kde, so we use the standard Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{2}$$

due to its convenient mathematical properties.

The bandwidth h is a free parameter which exhibits a strong influence on the resulting estimate. If Gaussian basis functions are used to approximate univariate data, and the underlying density being estimated is Gaussian then it can be shown that the optimal choice for h is

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{\frac{-1}{5}}, \tag{3}$$

where $\hat{\sigma}$ is the standard deviation of the samples [12].

Fig. 5 shows an example histogram and a pdf calculated using kde for two different cases of disturbances which occur in the cutting process.

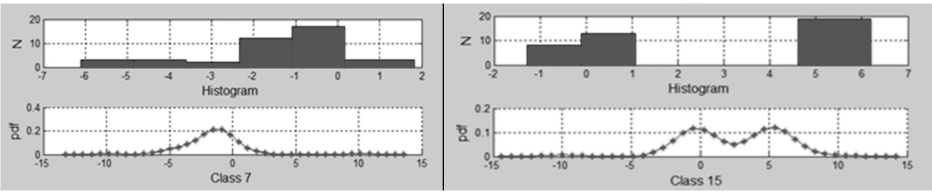


Fig. 5. Histograms of errors and pdfs for a machine which is out of order (classes 7 and 15)

3.3 Neural Network Classifier

A feedforward NN, Multi Layer Perceptron type [13], was used as a classifier. Such network can be considered as a universal approximator of any continuous function. The number of nodes in the input layer corresponds to the number of elements of the data vector, *i.e.*, the number of points, for which pdf values were appointed. Then the hidden layers exist, consisting of neurons characterized by nonlinear (log-sigmoid) activation function, and the output layer consisting of linear neurons. The number of neurons in the output layer depends on the assumed number of classes. This layer generates the final response of the network to a given pdf (target), which should equal 1 on that output which corresponds to the class from which the pdf comes from, and 0 for all other classes. Values between 0 and 1 are treated as a suggestion of belonging to a specific class (degree of truth).

A fundamental problem in modeling with NN is definition of network size and topology, *i.e.*, the number of hidden layers and choice of the number of neurons within individual layer. The number of input data elements is essential. For a small number of training samples and for a complex network structure a network overdimensioning may take place. Such network can adapt to the noise in the training data. If the number of samples is very high, the network cannot reproduce all training patterns. Estimation of the lower and higher range for Vapnik-Chervonenkis measure may help [14]. Usually, a general number of neurons is presupposed using heuristic principles, as a geometric mean of the number of inputs and outputs of the neural network, but their distribution into layers is performed empirically. Many different network structures were investigated, both one and two hidden layers, taking into account the effectiveness of choosing the correct class and the ability to generalization, understood as the ability to predict well beyond the training data. Best results were obtained for two hidden layers with 28 and 20 neurons respectively.

Learning is the ability to approximate the underlying behavior adaptively from the training data. The data is divided into two subsets: the training set $\mathcal{L}$ and the testing set $\mathcal{T}$. A validation subset $\mathcal{V}$ is allocated from set $\mathcal{L}$, used during the learning process, to estimate the degree of successful learning of the neural network [15]. Network's ability for generalization can be checked on the $\mathcal{T}$ set, on which the network was not trained. Data distribution into respective sets was: $\mathcal{L}$ - 80%, $\mathcal{V}$ - 10% and $\mathcal{T}$ - 10%.

During network training the Bayesian regulation backpropagation [16] and Levenberg-Marquardt (LM) [17] algorithms were tested. Both algorithms gave similar results in a form of the total percentage of correctly classified cases, but LM algorithm made it possible to obtain the results in much shorter time (lower number of epochs).

4 Testing the Classification Method

The proposed method of cutting error source classification was tested in respect of its ability to generalization, both in simulation based on adding noise to measurement data, and in using the test data coming from a real process.

Simulation data tests were performed to check the functioning of the network with the data obtained from the process, different from original data because of the existence of measurement errors. New data sets were generated by adding a Gaussian noise to a training sample. In this investigation the choice of noise variation characterized by SNR (signal to noise ratio) is important [15]. For SNR=20 the classifier revealed full repeatability of results obtained for original data.

Classification example for real data is shown for a machine, for which cutting errors exceeding the admissible values were observed. Classification results for the data obtained from improperly conducted process of cutting the board into pieces 760 mm long are presented in fig. 6. The histogram shows the occurrence of cutting deviations in the range from -3 to +5 mm with a „tail" and an asymmetry with respect to zero. The classifier classified the calculated pdf into class 11 with a degree of truth of 0.97.

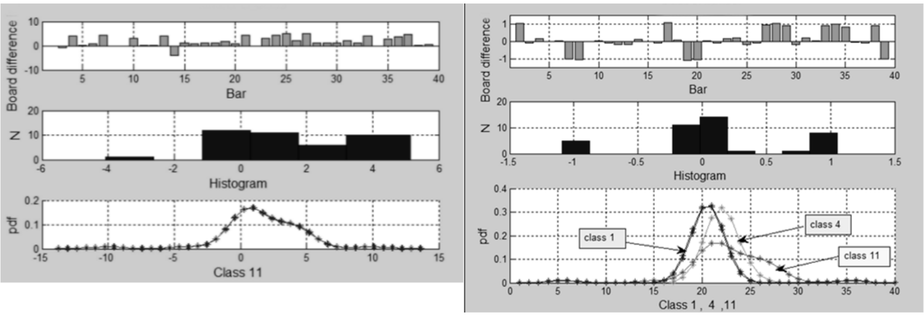


Fig. 6. Example data analysis for improperly conducted cutting process (left) and final result of classification (right)

This class shows the possibility of occurrence of 5 different faults in the process. The use of analysis employing bivariate distributions [2] revealed problems in the power transmission system of the cutoff. After correcting the faults, board measurements and classifier testing were performed again. This time the classifier indicated a dominant class 4 with a degree of truth of 0.83. This class shows, that improper measuring wheel calibration or paper slips may be the cause of errors. Because no slips were detected in the object, so the measuring wheel calibration was performed. After performing a test series of measurements, obtained results of classification were related to class 1 (*i.e.* correct course of the process) with a degree of truth of 0.98.

5 Conclusions

In the paper the results related to the use of NN-based classifier to assess causes of cutting errors in the rotary cutoff in the corrugated board machine processing line are presented. The investigation revealed, that the classifier of such design could recognize the correct class with the suggestion of degree of truth above 0.8. The tests confirmed high ability to generalize, NN network could recognize cases for data for which the procedures of network learning were not performed.

It should be emphasized, that the creation of such classifier requires sound knowledge and expert experience related to operation and service work of the machine, and analysis of data gathered for many different cases of incorrect machine operation. This knowledge is necessary for correct definition of classes describing causes of errors, and for connecting them to pdf of errors. Presented results are related to the first stage of investigation, aimed at confirming the hypothesis, that such approach in relation to the cutting process diagnostics is possible and efficient.

It appeared, that the created model of the classifier was able to recognize and assign the example cutting errors distribution to correct class of disorders.

Average time of calculations for complex network structures amounted about 2 hours on a PC. The tests revealed, that using different kernels does not influence substantially the classifier's results. The choice of appropriate *h* smoothing factor was more important.

The aim for further investigation is to refine the number of classes and causes of process disorders assigned to these classes. Especially it is about decreasing the number of symptoms assigned to a specific class, to make it possible for the user of the system to find quickly what corrective actions should be undertaken. As a result, a diagnostics system should be created, allowing for quick and reliable assessment of the cutoff performance and finding the source of possible errors.

References

1. Musielak, S.K.: Przekrawacz rotacyjny w tekturnicy. Część 1. Przegląd Papierniczy 1, 21–24 (2011)
2. Kasprzyk, J., Musielak, S.K.: Unconventional diagnostics of control system assessment for a cutoff used in a corrugated board machine process. In: Proc. of INPAP Conf. (2013)
3. Leonard, W.: Control of Electrical Drives. Springer, Berlin (2001)
4. Blechschmidt, J.: Taschenbuch der Papiertechnik. Carl Hanser Verlag, Műnchen (2010)
5. Larose, D.T.: Discovering Knowledge in Data. An Introduction to Data Mining. Wiley, New York (2005)
6. Barnaghi, P.M., Sahzabi, V.A., Bakar, A.A.: A Comparative Study for Various Methods of Classification. In: Int. Conference on Information and Computer Networks (ICICN 2012). IPCSIT, vol. 27, IACSIT Press, Singapore (2012)
7. Survey, N.: of Classification Techniques in Data Mining. In: Proc. of the Int. MultiConf. of Engineers and Computer Scientists IMECS 2009, Hong Kong, vol. 1 (2009)
8. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon, Oxford (1995)
9. Zhang, G.P.: Neural Networks for Classification: A Survey. IEEE Trans. on Systems, Man and Cybernetics Part C: Applications and Reviews 30(4) (2000)
10. Kulczycki, P.: Wykrywanie uszkodzeń w systemach zautomatyzowanych metodami statystycznymi. Wyd. Alfa. Warszawa (1998)
11. Wand, M.P., Jones, M.C.: Kernel smoothing. Chapman & Hall, New York (1995)
12. Silverman, B.W.: Density estimation for Statistics and Data Analysis. Chapman and Hall, New York (1986)
13. Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC (1961)
14. Vapnik, V.N., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16, 264–280 (1971)
15. Osowski, S.: Sieci neuronowe w ujęciu algorytmicznym. WNT, Warszawa (1996)
16. Cooper, G.E., Herskovits, E.: A Bayesian method for the induction of probabilistic networks for data. Machine Learning 9, 309–347 (1992)
17. Marquardt, D.W.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of the Soc. for Industrial and Applied Mathematics 11, 431–441 (1963)

RF Coverage and Pathloss Forecast Using Neural Network

Zia Nadir[1] and Muhammad Idrees Ahmad[2]

[1] ECE Dept., Sultan Qaboos University, PC. 123, P. Box. 33, Muscat, Oman
nadir@squ.edu.om
[2] DOMAS., Sultan Qaboos University, PC. 123, P. Box 36, Muscat, Oman

Abstract. The paper addresses the applicability of Okumura-Hata model in an area in Oman in GSM frequency band of 890-960 MHz. The Root Mean Square Error (RMSE) was calculated between measured Pathloss values and those predicated on the basis of Okumura-Hata model. We proposed the modification of model by investigating the variation in Pathloss between the measured and predicted values. This modification is necessary to consider the environmental conditions of OMAN. Artificial Neural Network (ANN) was also used to forecast the data for much larger distance. ANN provides a wide and rich class of reliable and powerful statistical tools to mimic complex nonlinear functional relationships. Here, feed forward Multilayer Perceptron (MLP) network was used. A typical MLP network consists of three layers i.e. input layer, hidden layer and output layer. The trained neural nets are finally used to make desired forecasts. These results are acceptable and can be used for OMAN.

Keywords: Pathloss model, Propagation models, Artificial Neural Network, Hata Model, Semi-Urban Area.

1 Introduction

In the design of any cellular mobile system, the fundamental task is to predict the coverage of the proposed system. Propagation models are useful for predicting signal attenuation or path loss which may be used as a controlling factor for system performance or coverage so as to achieve perfect reception [1]. It has been found that the mechanisms behind electromagnetic wave propagation are diverse and characterized by certain phenomena such as reflection, refraction and diffraction of waves. These phenomena induces signal scattering, fading and shadowing along the signal path and their effects can best be described (in a large scale) by the path loss exponent which defines the rate of change of attenuation that the signals suffers as it propagates from the transmitter to the receiver [2]. The wireless communication relies on the broadcast of waves in the free space. This also provides mobility for users and satisfies the demand of the customers at any location covered by the wireless network. Growth in the mobile communications field has now become slow, and has been linked to technological advancements [3,4]. The need for high quality and high capacity

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_36, © Springer International Publishing Switzerland 2014

networks, estimating coverage accurately has become extremely important. Therefore, for more accurate design coverage of modern cellular networks, signal strength measurements must be taken into consideration in order to provide an efficient and reliable coverage area.

This article addresses the evaluations between the statistical and the experimental analysis at GSM frequency of 900MHz. It was attained that, the most widely used propagation data for mobile communications is Okumura's measurements and this is recognized by the International Telecommunication Union (ITU) [5].

The cellular concept was a major breakthrough in solving the problem of spectral bottlenecks and user's capacity. It offered high capacity with a limited spectrum allocation without any major technological change. The cellular concept is a system level idea in which a single, high power transmitter is replaced with many low power transmitters. The area serviced by a transmitter is called a cell. Each small powered transmitter, also called a base station provides coverage to only a small portion of the service area. The power loss involved in transmission between the base station (BTS) and the mobile station (MS) is known as the Pathloss and depends particularly on the antenna height, carrier frequency, distance and environmental parameters. At higher frequencies the range for a given Pathloss is reduced, so more cells are required to cover a given area. Base stations close to one another are assigned different groups of channels. Neighboring base stations are assigned different groups of channels so that the interference between base stations or interaction between the cells is minimized. As the demand for service increases, the number of base stations may be increased, thereby providing additional capacity with no increase in radio spectrum. The key idea of modern cellular systems is that it is possible to serve the unlimited number of subscribers, distributed over an unlimited area, using only a limited number of channels, by efficient channel reuse [4]. The present models, discussed below have certain constraints. Accordingly these models need to be modified if we want these to be used for these regions where environmental and geographical differences are there.

2 Theoretical Propagation Models

Propagation models are mathematical representation of results of experiments conducted on the wave propagation under different frequencies, antenna heights and locations over different periods and distances. Propagation models indicate that average received signal power decreases logarithmically with distance [6]. They are divided into two basic types; namely: Free space propagation and Plane earth propagation model.

2.1 Free Space Propagation Model

In free space, the wave is not reflected or absorbed. Ideal propagation implies equal radiation in all directions from the radiating source and propagation to an infinite distance with no degradation. Spreading the power over greater areas causes the attenuation. Equation (1) illustrates how the power flux is calculated:

$$P_d = P_t \ / \ 4\pi d^2 \tag{1}$$

Where P_t is known as transmitted power (W/m^2) and P_d is the power at a distance d from antenna. If the radiating element is generating a fixed power and this power is spread over an ever-expanding sphere, the energy will be spread more thinly as the sphere expands.

2.2 Plane Earth Propagation Model

The free space propagation model does not consider the effects of propagation over ground. When a radio wave propagates over ground, some of the power will be reflected due to the presence of ground and then received by the receiver. Determining the effect of the reflected power, the free space propagation model is modified and referred to as the 'Plain-Earth' propagation model. This model better represents the true characteristics of radio wave propagation over ground. The plane earth model computes the received signal to be the sum of a direct signal and that reflected from a flat, smooth earth. The relevant input parameters include the antenna heights, the length of the path, the operating frequency and the reflection coefficient of the earth. This coefficient will vary according to the terrain type (e.g. water, desert, wet ground etc). Pathloss Equation for the plane Earth Model is illustrated in equation (2).

$$L_{pe} = 40\log_{10}(d) - 20\log_{10}(h_1) - 20\log_{10}(h_2) \tag{2}$$

Where d represents the path length (m) and h_1 and h_2 are the antenna heights (m) at the base station and the mobile, respectively. The plane earth model in not appropriate for mobile GSM systems as it does not consider the reflections from buildings, multiple propagation or diffraction effects. Furthermore, if the mobile height changes (as it will in practice) then the predicted Pathloss will also be changed.

3 Empirical Propagation Models

Empirical propagation models will be discussed in this section; among them are Okumura and Hata models.

3.1 Cellular Propagation Models

The two basic propagation models (free space loss and plane-earth loss) would require detailed knowledge of the location, dimension and constitutive parameters of every tree, building, and terrain feature in the area to be covered. This is far too complex to be practical and would yield an unnecessary amount of detail. One appropriate way of accounting for these complex effects is via an empirical model. There are various empirical prediction models among them are, Okumura – Hata model, Cost 231 – Hata model, Cost 231 Walfisch – Ikegami model, Sakagami- Kuboi model. These models depend on location, frequency range and clutter type such as urban, sub-urban and countryside.

3.2 Okumura's Measurements

Okumura carried out extensive drive test measurements with range of clutter type, frequency, transmitter height, and transmitter power. It states that, the signal strength decreases at much greater rate with distance than that predicted by free space loss [5, 7-8].

3.3 Hata's Propagation Model

Hata model was based on Okumura's field test results and predicted various equations for Pathloss with different types of clutter. It is well suited model for the Ultra High Frequency (UHF) band [9]. The limitations on Hata Model due to range of test results from carrier frequency *150MHz to 1500MHz*, the distance from the base station ranges from 1Km to 20Km, the height of base station antenna (h_b) ranges from 30m to 200m and the height of mobile antenna (h_m) ranges from 1m to 10m. It was also observed that the signal strength is a function of distance and antenna height, as we can see in this work the highest antenna has less propagation path loss and as the distance increases the path loss also increases [10]. Hata created a number of representative Pathloss mathematical models for each of the urban, suburban and open country environments, as illustrated in following equations, respectively. Okumura takes urban areas as a reference and applies correction factors as following:

Urban areas: $L_{dB} = A + B \log_{10} R - E_{1,2.3}$

Suburban areas: $L_{dB} = A + B \log_{10} R - C$

Open areas: $L_{dB} = A + B \log_{10} R - D$

Where

$A = 69.55 + 26.16 \log_{10} fc - 13.82 \log_{10} h_b$

$B = 44.9 - 6.55 \log_{10} h_b$

$C = 2 (\log_{10} (f_c / 28))^2 + 5.4$

$D = 4.78 (\log_{10} f_c)^2 + 18.33 \log_{10} f_c + 40.94$

$E_1 = 3.2 (\log_{10} (11.7554\, h_m))^2 - 4.97$

for large cities, $fc \geq 300$MHz.

$E_2 = 8.29 (\log_{10} (1.54\, h_m))^2 - 1.1$

for large cities, $fc < 300$MHz.

$E_3 = (1.1 \log_{10} f_c - 0.7)\, h_m - (1.56 \log_{10} f_c - 0.8)$

for medium to small cities.

Definition of Parameters:

h_m ;mobile station antenna height [m]

d_m ;distance between the mobile and the building [m]

h_o ; typical height of a building above local terrain height [m]

h_b ;base station antenna height above local terrain height [m]

r; great circle distance between base station and mobile [m]

$R=r \times 10^{-3}$ great circle distance between BS and mobile [km]

f ;carrier frequency [Hz]

$fc=f \times 10^{-6}$ carrier frequency [MHz]

λ ; free space wavelength [m].

The practical Pathloss can be calculated using the equation:

$$L_P\,(dB) = P_t - P_r$$

Where P_t is the transmitted power which is equal to 47dB and P_r is the received power. Whereas, the Pathloss for Okumura-Hata Model can be calculated by the following equation:

$A= 69.55+26.16 \log10 f_c -13.82 \log10 h_b$, and $E_3= (1.1 \log10 f_c – 0.7) h_m – (1.56 \log10 f_c – 0.8)$ for small and medium city. [11-14]

The generation of such measurements is based on the assumption that the power of a signal decreases monotonically with the increase of the distance traveled by the signal [15]. Thus, Hata model is not suitable for micro-cell planning where antenna is below roof height and its maximum carrier frequency is 1500MHz. It is not valid for 1800 MHz and 1900 MHz systems.

4 Results and Discussions

To generate measurements of signal strength level for downlink and uplink at coverage areas for a cell, TEMS tools were used. However, the road of Al Khuwair can be considered as an urban area of Okumura-Hata model was used. For this paper, experimental data set (named as set-B) is used.

After determining the Pathloss of the practical measurements for each distance, the study was carried on in order to make a comparison between the experimental and theoretical values and the result is shown in Fig-1.

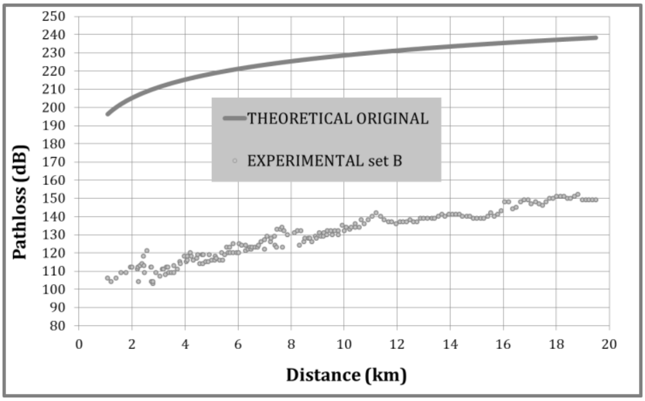


Fig. 1. Theoretical and Experimental Pathloss vs. distance Set-B [14]

From the above plots, the results clearly show that the measured Pathloss is less than the predicted Pathloss by a difference varying from 4 to 20dB. However, there are several reasons which may cause those significant differences. First of all, in Japan there are few areas virtually satisfying the conditions; and if any, they are narrow. Moreover, the geographical situation of Japan is different from that in Oman due to geographical differences. Then, Root Mean Square Error (*RMSE*) was calculated between measured Pathloss value and those predicted by Hata model using the following equation [16-17]:

$$\mathrm{RMSE} = \sqrt{\left(\sum \frac{(\mathrm{P_m}-\mathrm{P_r})^2}{(\mathrm{N}-1)}\right)} \tag{3}$$

Where;

P_m: Measured Pathloss (dB); P_r: Predicted Pathloss (dB) ; N: Number of Measured Data Points

The *RMSE* was found greater than 110dB but the acceptable range is up to 6 dB [17]. Therefore, the *RMSE* is adjusted with the Hata equation for urban area and the modified equation will be as following:

$$\begin{aligned} &\mathrm{L_{p(modified)}(Urban)} = 69.55 + 26.16\log_{10}(f) - 13.82\log_{10}(h_b) \\ &+(44.9 - 6.55\log_{10}(h_b))\log_{10}(d) \pm \mathrm{MSE} \\ &-(1.1\log_{10}(f) - 0.7)h_m - (1.56\log_{10}(f) - 0.8) \end{aligned} \tag{4}$$

The modified result of Hata equation is shown in Fig. 2 and the *RMSE* in this case is less than 6dB, which is acceptable [14].

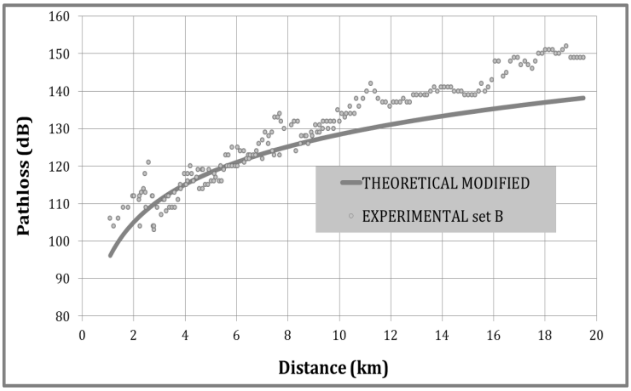


Fig. 2. Modified Theoretical and Experimental Pathloss–Set B

In order to verify that the modified Hata's equation (4) is applicable for other areas of Oman, another data generated from TEMS tool for another cell in the road of Al Khuwair has been used. Based on that practical data, the propagation Pathloss and the distance have been re-verified for another cell [11] but not shown in this current paper. However, on experimental Set B, few data points are a bit far from interpolated values which can be due to the nature of cell B with high rise buildings.

Forecasting Using Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) provides a wide and rich class of reliable and powerful statistical tools to mimic complex nonlinear functional relationships. In this work we used feed forward Multilayer Perceptron (MLP) network. A typical MLP network consists of three layers i. e. input layer, hidden layer and output layer. Each of these layers may comprise of several neurons and synapses which are connected to each other through a predesigned structure. These are usually represented by directed graphs containing vertices, edges and nodes. The networks are then trained by learning through empirical data. These trained neural nets are finally used to make desired forecasts. The design of the architecture and the training process of the networks require the choice of the scaling method, number and form of the input nodes, number of hidden neurons, the activation function, an error function and the type of output layer. We used neuralnet package of R statistical computing [19] for the training of the network. This package uses back propagation algorithm for adaptive learning of the neural network. The logistic activation function was chosen in the hidden layers and the linear output type was selected. The topology of the trained network along with synaptic weights is presented in Fig.3. The input layer consists of tree neurons. Here U is the scaled distance by 1000 and the other two neurons were decided to be the sine and cosine functions of distance. Several different sets of nodes for the hidden layer were initially considered and compared based on the error function and the number of steps takes to train the network. Finally three hidden neurons gave the best performance with minimum errors of 0.07 and 216 steps to train the network. This trained network was then used to predict the path losses up to a distance of about three kilometers. The path losses predicted by this neural network are plotted along with those obtained through theoretical model and experimental data in Fig.4. This graph clearly reveals that the losses generated by the neural network are intuitively more appealing and much closer to the experimental data. This vindicates power and ability of the neural network to simulate complex phenomenon.

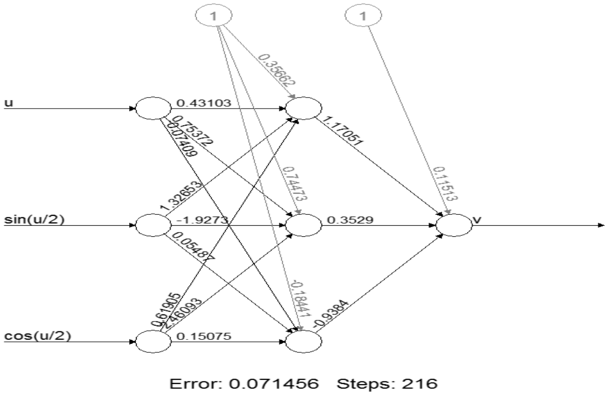


Fig. 3. Architecture of the final Neural Network Used

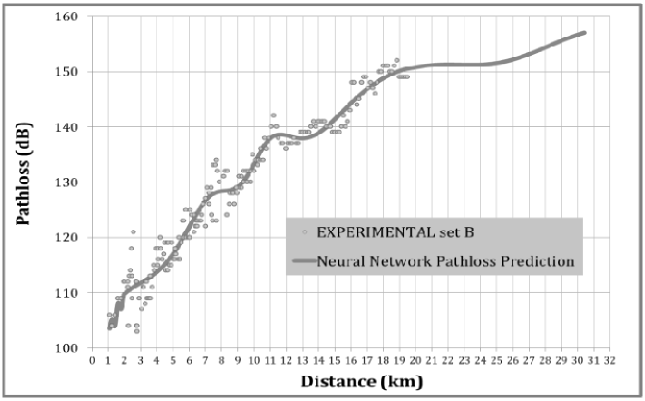


Fig. 4. Pathloss versus distance for Experimental set-B, Theoretical, and Neural Network Forecast

5 Conclusion

This work was focused for predicting the root mean signal strength in different areas. However, most propagation models aim to predict the median Pathloss. But, existing predictions models differ in their applicability over different terrain and environmental conditions. Although there are many predictions methods based on deterministic processes through the availability of improved databases, but the Okumura-Hata model is still mostly used [17-18]. That is because of the ITU-R recommendation for its simplicity and its proven reliability.

The effects of terrain situation predicted at 900MHz were analyzed. Results of radio signal propagation measurements for an urban area and extended to semi urban area for ANN forecast in Oman was compared to those predicted based on Okumura-Hata model. However, the Okumura-Hata propagation model might not be fully adapted in Oman because there is no rain attenuation impact in Oman environment due to lack of rain. Therefore, further improvement of Okumura-Hata model in the urban area has been suggested. This improvement was achieved by using root mean square error (*RMSE*) between measured and predicted Pathloss values in order to provide sufficient *RMSE* for radio prediction.

The area under investigation can be treated as a combination of urban and semi-urban area. Also, if more detailed environmental information is included in the model, better prediction results might be achieved. ANN gave us a reasonable forecast of Pathloss for a relative larger area. Also the missing experimental points show a good agreement with Theoretical model.

Acknowledgement. The authors would like to thank SQU for the support to carry out this study.

References

1. Shveta, S., Uppal, R.S.: RF Coverage Estimation of Cellular Mobile Systems. International J. of Engineering and Technology 3(6), 398–403 (2012)
2. Ubom, E.A., Idigo, V.E., Azubogu, A.C.O., Ohaneme, C.O., Alumona, T.L.: Path loss Characterization of Wireless Propagation for South – South Region of Nigeria. International J. of Computer Theory and Engineering 3(3), 360–364 (2011)
3. Nobel, D.: The history of land to mobile radio communications. IEEE Transactions on Vehicular Technology, 1406–1416 (1962)
4. MacDonald, V.H.: The cellular concept. The Bell Systems Technical Journal 58(1), 15–43 (1979)
5. Medeisis, A., Kajackas, A.: On the Use of the Universal Okumura-Hata Propagation Predication Model in Rural Areas. In: Vehicular Technology Conference Proceedings, vol. 3, pp. 1815–1818. VTC Tokyo, Japan (2000)
6. Manju, K., Tilotma, Y., Pooja, Y.: Comparative Study of Path Loss Models in Different Environments. International J. of Engineering Science and Technology (IJEST) 3(4), 2945–2949 (2011) ISSN: 0975-5462
7. Hata, M.: Empirical Formula for Propagation Loss in Land Mobile Radio Services. IEEE Transactions on Vehicular Technology VT 29(3), 317–326 (1980)
8. Wilson, R.D., Scholtz, R.A.: Comparison of CDMA and Modulation Schemes for UWB Radio in a Multipath Environment. In: Proceedings of IEEE Global Telecommunications Conference, vol. 2, pp. 754–758 (2003)
9. Rakesh, N., Srivatsa, S.K.: A Study on Pathloss analysis for GSM Mobile Networks for Urban, Rural and Suburban Regions of Karnataka State. International J. of Distributed and Parallel Systems (IJDPS) 4(1), 53–66 (2013), doi:10.5121/ijdps.2013.4105

10. Shalangwa, D.A., Singh, S.K.: Measurement and Modeling of Path Loss for GSM900 in SubUrban Environment over Irregular Terrain. International Journal of Computer Science and Network Security (IJCSNS) 10(8), 268–274 (2010)
11. Zia, N., Nazar, E., Farid, T.: Pathloss determination using Okumura-Hata model and spline interpolation for missing data for Oman. In: World Congress on Engineering, IAENG-WCE-2008. Imperial College, London (2008)
12. Zia, N., Mohammad, I.A.: Pathloss Determination using Okumura-Hata Model and Cubic regression for missing Data for Oman. In: International Conference on Communications Systems and Applications, IAENG-ICCSA 2010, Hong Kong (2010)
13. Zia, N., Mohammad, I.A.: Characterization of Pathloss using Okumura-Hata Model and missing Data Prediction for Oman. IAENG Transactions on Engineering Tech., vol. (5): Special Edition of the Int. Multi-conference of Engineers and Computer Scientists 2009. AIP Conference Proceeding, vol. 1285, pp: 509–518 (2010)
14. Zia, N.: Empirical Pathloss Characterization for Oman. In: IEEE Computing, Communications & Applications Conference 2012 (IEEE ComComAP 2012). HongKong University of Science and Technology, Hong Kong (2012)
15. Mourad, F., Snoussi, H., Kieffer, M., Richard, C.: Robust bounded-error tracking in wireless sensor networks. In: SYSID 2012, Brussels, Belgium (2012)
16. Shalangwa, D.A., Jerome, G.: Path Loss Propagation Model for Gombi Town Adamawa State Nigeria. International J. of Computer Science and Network Security (IJCSNS) 10(6), 186–190 (2010)
17. Wu, J., Yuan, D.: Propagation Measurements and Modeling in Jinan City. In: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Boston, MA, USA, vol. 3, pp. 1157–1159 (1998)
18. Okumura, Y., et al.: Field Strength and Its Variability in VHF and UHF Land-Mobile Radio Service. Review of the Electrical Communications Laboratory 16(9-10) (1968)
19. Frauke, G., Stefan, F.: neuralnet: Training of Neural Networks. R. Journal 2/1 (2010) ISSN 2073-4859

Automatic Generator Re-dispatch for a Dynamic Power System by Using an Artificial Neural Network Topology

Ahmed N. AL-Masri

Management and Science University, Faculty of Information Science and Engineering,
Department of Engineering and Technology, 40100 Shah Alam, Selangor, Malaysia
ahmed_naufal@msu.edu.my

Abstract. In recent years, the number of severe fault situations and blackouts worldwide has increased with the growth of large interconnected power system networks. This paper attempts to investigate the effect of a contingency (N-1) on rotor angle stability and thermal line flow for a dynamic power system. In addition, a solution is presented to eliminate system instability by providing an automatic generator re-dispatch instantly after a disturbance. Based on the ability of an Artificial Neural Network (ANN), it is possible to model a mathematical relationship between a power system disturbance and a control action due to the fast response of an ANN system. This relationship is obtained by the neurons between the input and output layers of the ANN topology. The completed model and data knowledge preparation process were successfully tested on an IEEE 9-bus test system. The ANN was able to provide a control action in a very short time period with high accuracy. An optimal amount of generator re-dispatch in Megawatt (MW) can contribute towards eliminating bus voltage and thermal line flow violations or unstable power system operation.

Keywords: Dynamic Power System, Rotor Angle Stability, Generator Re-dispatch, Artificial Neural Network, Back propagation algorithm.

1 Introduction

With open access to deregulated markets and high economic growth, the consequent electrical power transfers are forcing the transmission systems to run at their limit. As a result, unexpected events such as weak interconnections, high loading of lines and protection failures may cause the systems to lose security and that increases the possibility of catastrophic failures or blackouts. In recent years the number of blackouts and the associated negative consequences has grown. Analyzing these catastrophes shows that the operating guidelines used for many years were based on offline stability studies, which tend to be conservative for normal conditions and inaccurate for unexpected unusual events.

In many cases a static study cannot achieve the necessary stability under changing grid and generation conditions [1]. Worldwide major grid blackouts were investigated and discussed by the engineers attending the IEEE power Engineering Society

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_37, © Springer International Publishing Switzerland 2014

General meeting in Denver, Colorado in 2004 [2]. Furthermore, the committee provided a summary of blackouts in North America and Europe over 2003 and the causes of each blackout were described by Andersson et al.[3].

It is difficult to follow the changes in the status of a dynamic power system by using only offline analysis. Therefore, online analysis is used to obtain a more accurate picture and to reflect the actual behaviour of the power system [4]. Online simulation using actual operating conditions reduces uncertainty, and can be used for control adaptation [5]. The challenging part of a practical power system is the possibility of predicting the dynamic behaviour of the power system and to provide a solution when a disturbance occurs. The solution could be offered by an intelligent system and the application of static stability analysis results can be used to provide a remedial action.

As reported by a CIGRE research group [6], study is needed in the implementation of automated control action. That is to say, systems in which Dynamic Security Assessment (DSA) tools make security assessments, determine remedial measures, and automatically take absolute actions. Such systems must be adaptive in order to respond to system changes and require a high degree of robustness and reliability. For the control action, the DSA must offer the following assets:

- Ranking of voltage control supplies (generator or shunt compensation) to correct voltage violations.
- Suggested active power for re-dispatch (amount) to alleviate thermal limit violations and avoid angular instability.
- Suggested MW re-dispatch (location) to improve damping.
- Suggested load shedding (amount and location) to move from alert/emergency states to a secure state.

The main objectives of this study are: first, to take control action based on a dynamic analysis result in order to bring the system from a critical unstable state to a stable region. If successful, this could lead to the alert state, where actions may be necessary to achieve the normal state, or it could lead directly to the normal state. Second, a data processing algorithm is presented to solve the data knowledge preparation problem in an Artificial Intelligence (AI) model and keep the system updated with any changes in the power system. Moreover, a snapshot of the actual dynamic system condition is taken and a comprehensive stability analysis with sufficient speed is proceeded to allow the operator to take preventive actions to ensure stable operation is resumed.

The neural network algorithm has been implemented in this application because of its performance in predication of the optimal value of watt that is necessary to maintain the power system stability. Whereas the other algorithms such as Gaussian processes, Markov models and Fuzzy logic also can be implemented especially for control system. However, the neural network performance is better than other methods in term of prediction. Therefore, the neural network in this paper is selected to predict the generator value which can be used to ensure the system stability instantly after disturbance.

2 Artificial Neural Network (ANN) Topology for a Power System Model

The application of Artificial Neural Networks in power systems is already prevalent in many applications. Neural Network architectures have been classified into several types based on their learning mechanisms [7]. Most of the intensive classification of ANN models are used to solve different types of mathematical problems [8]. Back propagation is one of the simplest and general learning algorithms and it is more instructive than the other methods, which gives greater advantages for power system stability. In this current study, a developed feed forward back propagation algorithm is used to solve the power control (generation dispatch/ load shedding) problem due to changes in the aspects of contingencies.

The methodology of the work is presented in Figure 1. The algorithm is divided into three major parts:

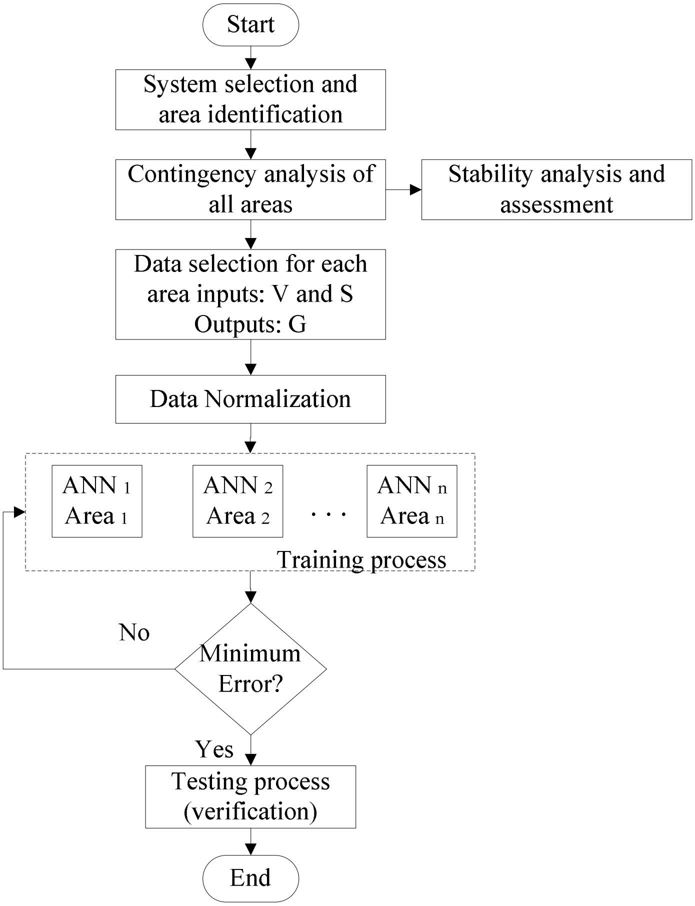

Start
System selection and area identification
Contingency analysis of all areas
Stability analysis and assessment
Data selection for each area inputs: V and S Outputs: G
Data Normalization
ANN 1 Area 1
ANN 2 Area 2
. . .
ANN n Area n
Training process
No
Minimum Error?
Yes
Testing process (verification)
End

Fig. 1. ANN system methodology

A. Contingency and stability analyses

This stage of analysis is required before proceeding to the data knowledge preparation model. However, the system is selected based on a region or area for reducing the number of inputs and patterns, which would speed up the ANN processing time for training and testing.

Contingency analysis allows the operator to examine the system under different operational conditions within the system criteria. Eventually, the design engineers are required to maintain secure system operation within the system criteria based on the test results. In addition, the operator will have the ability to deal with most contingency events such as line outage and a generation trip for (N-1) contingencies. The operator must act quickly before cascading failure occurs, which may cause a system blackout or separate the system into islands. In the proposed algorithm, by using contingency analysis data supported by the historical data of the system operation, the ANN is able to provide the optimal amount of generator re-dispatch under different contingency cases. These parameters are based on a model of the power system that is used to study the outage events and make an automatic decision rather than alert the operators to overloads or voltage limit violations. Contingency events correspond to changes in network admittances. As a result, the network reconfiguration can be estimated using the sensitivities of voltages, reactive outputs and thermal flows with respect to the admittance changes [9].

B. ANN model

The ANN is designed to be the second step of the developed algorithm when the system is analyzed and the assessment report is saved for record. However, the data knowledge is prepared in a static state and for each type of contingency it is declared whether any generator re-dispatch is required or otherwise. Subsequently this data is used to train the ANN after the normalization process. The normalization process is to make all the data in the range -1 to +1. The neurons are able to deal with this range since the tangent function was selected to be the neuron function as given in Equation (1) and following Figure 2.

$$O_k = f\left(net_k\right) = \frac{e^{\lambda net_k} - e^{-\lambda net_k}}{e^{\lambda net_k} + e^{-\lambda net_k}} \tag{1}$$

where, O_k is the neuron for layer k and λ determines the shape of the function = 1 for the network k.

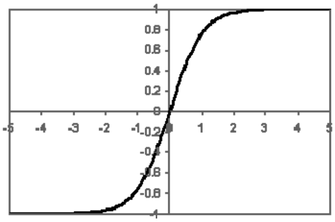


Fig. 2. Tangent function for λ =1

The back propagation technique is considered as a suitable learning method due to the design of the dynamic power system to be based on the historical data knowledge from the static stage. The back propagation method gives the ability to adapt the weights when new inputs are introduced. Finally, the error is calculated during every single iteration and the learning procedure repeated for all patterns ($p = 1,2,..., N$) or epochs ([Input, Output]) until the specified threshold value of error is reached or until a total iteration is reached. Although the back propagation learning algorithm has a highly mathematical foundation, it shows better accuracy than other methods especially when a small space gap in load scale is used as will be described in the Simulation result section.

C. Testing process (verification)

An optimal amount of generator re-dispatch in MW should be suggested for each unstable operation case. This amount is the output of the ANN when untrained data is used as input for the same system or area that has been trained. Depending on the selection of the system if divided into a number of areas, the generator can be re-dispatched locally or for the whole system. For example, if the system were to be divided into two areas, therefore, the contingency and stability analysis will be for each area individually and similarly the output of the ANN would be for each area as well.

3 Simulation Result

The algorithm was developed to be used for any dynamic power system. In this paper, a simple 9-bus test system was used for demonstrating the ability of using the ANN for generation re-dispatch. The system was as shown in Figure 3 consisting of nine buses, six transmission lines, three generators and three loads. The percentage rate of the transmission lines are based on their MVA rating, which represents the line limits. In the base case, all the loads were supported by the amount of generation in a stable operation. The system can stand some contingencies but only up to the point when a heavy line is disconnected. The ANN is applied to support the power system at normal operation when no control action is required and when the system is under large disturbance as well. The line overloading and bus voltage violation cases were monitored by simulating a single contingency (N-1) under varying load conditions. This data is considered as the pattern (p) for the ANN cases. Furthermore, the thermal line flow and bus voltage were calculated using a steady state Newton Raphson load flow method.

The steady state control action implementation including voltage and transmission line flow violations have been discussed in detail in a previous paper [10]. Moreover,

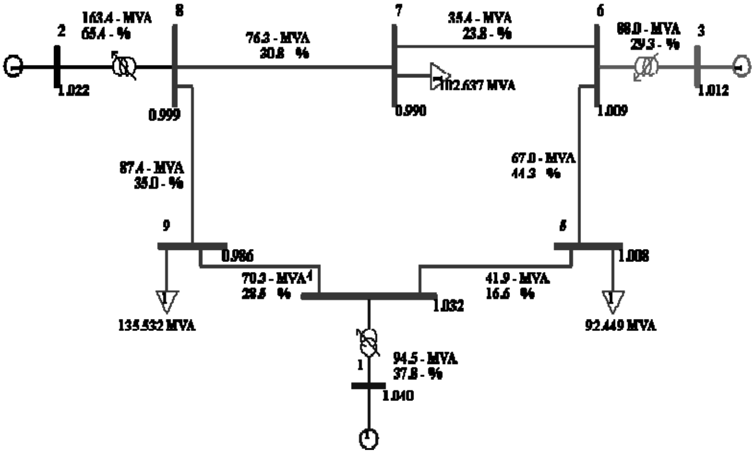
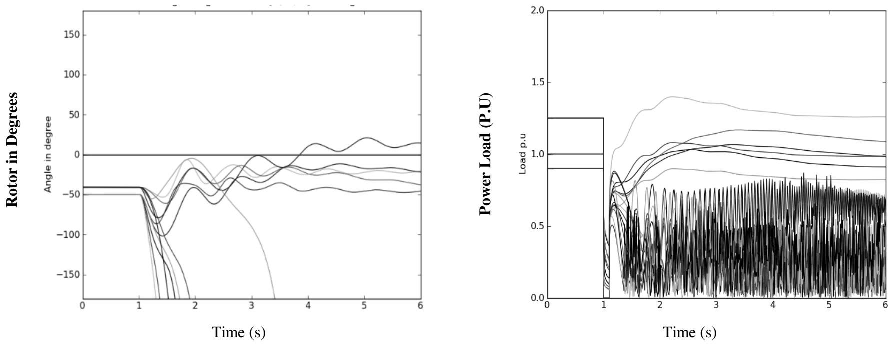


Fig. 3. Single line diagram for 9-bus test system

the dynamic power system stability assessment and rotor angle stability analysis for the same system and the 87-Malaysia power system were reported in a previous journal [11].

The rotor angle stability was determined by comparing the rotor angle difference between any generators to a reference generator whereby the rotor angle of the first generator was considered as the reference angle. Unstable operation occurred when at least one generator lost its synchronism with the other generators or the rotor angle (δ) was beyond its operational range of -180°< δ < 180° as shown in Figure 4 (a). As it can be seen in Figure 4 (b), the load has lost its stability and should be disconnected from the main grid immediately after losing synchronism between rotor angles.

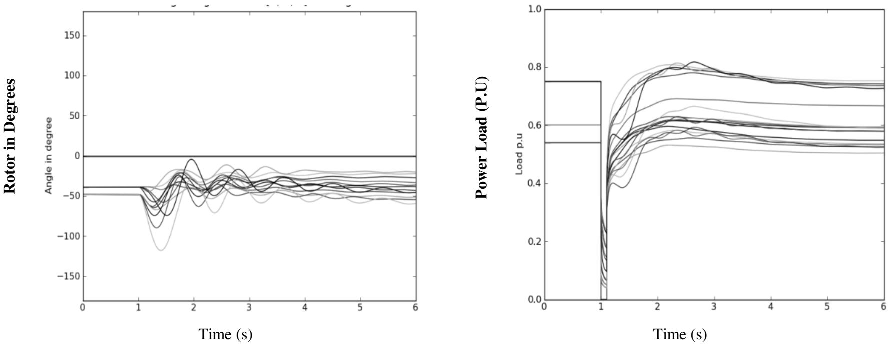


(a) Rotor angle stability.

(b) Load power in P.U.

Fig. 4. Stability assessment at full load level

Control action is highly desirable for the system to remain in stable operation when the system suffers an unstable condition. When the total load level reduced to 61 % constant division to all loads, the system was running in a stable condition even when the disturbance occurred and cleared after 100 ms as shown in Figure 5. However, the system does not require any generators to be re-dispatched (the ANN output is zero).



(a) Rotor angle stability.

(b) Load power in P.U.

Fig. 5. System stability at low load level

The ANN parameters for the training process were selected as given in table 1. The reason of each value selection was explained in details by reference [11].

Table 1. The ANN parameters values

Parameter	Value
Performance	0.01
Epochs	2000
Learning rate	0.01
Momentum rate	0.1
Hidden neurons	12

The performance of the Root Mean Square Error (RMSE) formula attained a value of 0.04116. The Momentum factor and learning rate were 0.1 and 0.01 respectively. The learning rate was chosen to be low which led to a slow learning process so as to increase the network performance.

The developed algorithm was verified by using a test data set that had not been trained before. The ANN model was able to predict the optimal amount of generation re-dispatch that was required is each case to maintain system stability. The total number of patterns for testing and training data is presented in table 2.

The training data was generated based on the load level, starting from minimum operation condition at 60 % up to the maximum load level at 100 %. Each load level contains seven contingencies including the base case. The alert status was monitored for any more generation re-dispatch cases observed when the system was loaded over 80 %.

The correlation coefficient between the PSS™E (Power System Simulation for Engineering, which is developed by Siemens) simulation and the ANN results was tested. Table 2 includes the accuracy of the ANN for each generator and load for all the test data.

Table 2. Correlation coefficient for different testing load scenarios

	Total patterns number	Load levels (+2 load scale)	
Training data	147	60% to 100%	
Testing data	140	61% to 99%	
Correlation coefficient	G_1	G_2	G_3
for all testing data	1	0.97527	0.972817

The generator re-dispatch schemes are illustrated in Figure 6 to Figure 8 for each contingency case. The value of the generator re-dispatch was measured to be a high value at a high load level and low at for a lower load level. However, these amounts depended on generator reaction among disturbance values and location.

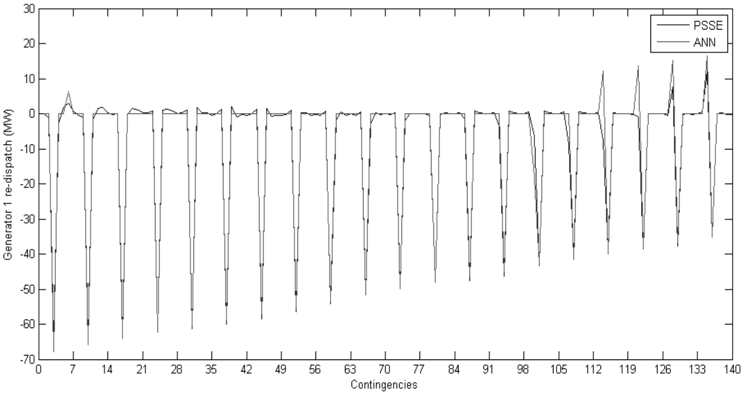


Fig. 6. Generator 1 re-dispatch control scheme

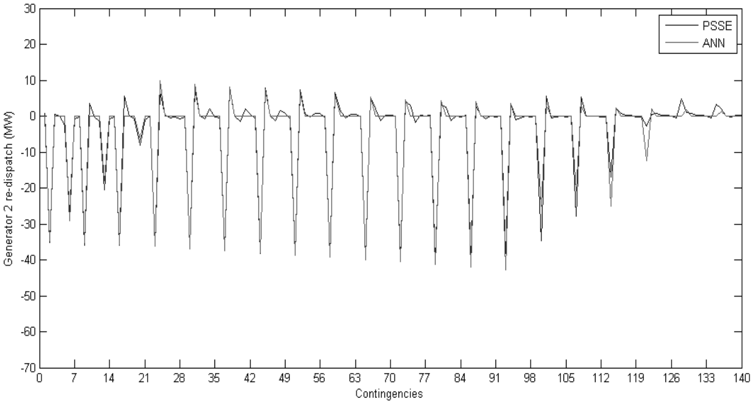


Fig. 7. Generator 2 re-dispatch control scheme

The ANN time response of the testing process was less than one millisecond for each contingency, which gave a good result for the dynamic power system and suggested control action.

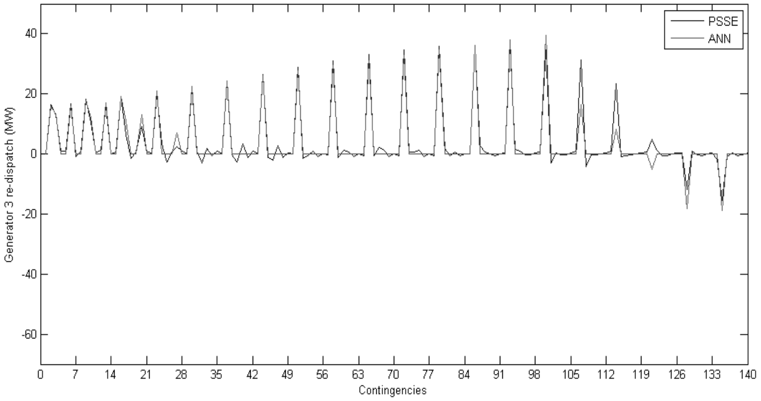


Fig. 8. Generator 3 re-dispatch control scheme

4 Conclusion

This paper has presented an automatic generator re-dispatch scheme for a dynamic power system by using an ANN technique. The effect of a contingency (N-1) on the dynamic power system stability was investigated. A back propagation algorithm for remedial action was developed in this paper to give a suggested value of generator re-dispatch in MW. The dynamic data (rotor angle, load level and contingency type) was used in combination with the static data (generator re-dispatch) as inputs and outputs for the ANN, respectively. Based on the results, the neural network provided a corrective action based on the stability criteria of the power system at that particular load level. One of the more significant findings to emerge from this study is that the values of generation re-dispatch are virtually instantaneously and accurately estimated by the ANN.

5 Future Work and Recommendation

The presented work is considered a starting point for an achievable fully automated dynamic power system control. In order to complete this work, several recommendations are suggested:

- The same algorithm should be tested on a large power system with a heavy interconnected load.
- Another artificial intelligence technique should be considered due to the limitation of using ANN technology. One of the ANN drawbacks is the high requirement of a design model and data processing.

- Load shedding could be considered as an emergency control action when a heavy disturbance hits the power system or a heavy transmission line outage occurs.

A real time system should be considered as practical work within this research and system verification also should also be demonstrated.

References

1. Lerch, E., Ruhle, O.: Dynamic Security Assessment to protect systems after severe fault situations. In: International Conference on Power System Technology (PowerCon), Chongqing, pp. 1–6 (2006)
2. Kolluri, S., He, T.: Design and operating experience with fast acting load shedding scheme in the Entergy system to prevent voltage collapse. In: IEEE Power Engineering Society General Meeting, pp. 1625–1630 (2004)
3. Andersson, G., Donalek, P., Farmer, R., Hatziargyriou, N., Kamwa, I., Kundur, P., Martins, N., Paserba, J., Pourbeik, P., Sanchez-Gasca, J.: Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance. IEEE Transactions on Power Systems 20, 1922–1928 (2005)
4. Balu, N., Bertram, T., Bose, A., Brandwajn, V., Cauley, G., Curtice, D., Fouad, A., Fink, L., Lauby, M., Wollenberg, B.: On-line Power System Security Analysis. Proceedings of the IEEE 80, 262–282 (1992)
5. Grigsby, L.: Power System Stability and Control. CRC Press (2007)
6. Brochure, C.T.: Review of On-line Dynamic Security Assessment Tools and Techniques. Technical report, CIGRE Working Group C4.601 (2007)
7. Rajasekaran, S., Pai, G.: Neural Networks, Fuzzy Logic and Genetic Algorithms: Synthesis and Applications. PHI Learning Pvt. Ltd. (2004)
8. Vankayala, V., Rao, N.: Artificial Neural Networks and Their Applications to Power Systems-a Bibliographical Survey. Electric Power Systems Research 28, 67–79 (1993)
9. Ruiz, P., Sauer, P.: Voltage and Reactive Power Estimation for Contingency Analysis Using Sensitivities. IEEE Transactions on Power Systems 22, 639–647 (2007)
10. Al-Masri, A., Kadir, M.A., Hizam, H., Mariun, N., Yusof, S.: Control Action Based on Steady-State Security Assessment Using an Artificial Neural Network. In: IEEE International Conference on Power and Energy (PECon), Kuala Lumpur, pp. 706–711 (2010)
11. Al-Masri, A., Kadir, M.A., Hizam, H., Mariun, N.: A Novel Implementation for Generator Rotor Angle Stability Prediction Using an Adaptive Artificial Neural Network Application for Dynamic Security Assessment. IEEE Transactions on Power Systems, 1–10 (2013)

A New Bat Based Back-Propagation (BAT-BP) Algorithm

Nazri Mohd. Nawi, Muhammad Zubair Rehman, and Abdullah Khan

Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia (UTHM),
P.O. Box 101, 86400 Parit Raja, Batu Pahat, Johor Darul Takzim, Malaysia
nazri@uthm.edu.my, zrehman862060@gmail.com,
hi100010@siswa.uthm.edu.my

Abstract. Metaheuristic algorithm such as BAT algorithm is becoming a popular method in solving many hard optimization problems. This paper investigates the use of Bat algorithm in combination with Back-propagation neural network (BPNN) algorithm to solve the local minima problem in gradient descent trajectory and to increase the convergence rate. The performance of the proposed Bat based Back-Propagation (Bat-BP) algorithm is compared with Artificial Bee Colony using BPNN algorithm (ABC-BP) and simple BPNN algorithm. Specifically, OR and XOR datasets are used for training the network. The simulation results show that the computational efficiency of BPNN training process is highly enhanced when combined with BAT algorithm.

Keywords: Back propagation neural network, bat search algorithm, local minima, artificial bee colony algorithm.

1 Introduction

Artificial Neural Networks (ANNs) are diagnostic techniques sculpted on the learning and neurological functions of the human brain. ANNs works by processing information like biological neurons in the brain and consists of small processing units known as Artificial Neurons, which can be trained to perform complex calculations [1].

An Artificial Neuron can be trained to store, recognize, estimate and adapt to new patterns without having the prior information of the function it receives. This ability of learning and adaptation has made ANN superior to the conventional methods used in the past. Due to its ability to solve complex time critical problems, it has been widely used in the engineering fields such as biological modeling, financial / weather forecasting, decision modeling, control systems, manufacturing, health and medicine, ocean and space exploration, and noise-induced hearing loss (NIHL) etc. [2 - 9]

An Artificial Neural Network (ANN) consists of an input layer, one or more hidden layers and an output layer of neurons. In ANN, every node in a layer is connected to every other node in the adjacent layer. ANN are usually classified into several categories on the basis of supervised and unsupervised learning methods and feed-forward and feed-backward architectures [1]. Back-Propagation Neural Network

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_38, © Springer International Publishing Switzerland 2014

(BPNN) algorithm is the most popular and the oldest supervised learning multilayer feed-forward neural network algorithm proposed by Rumelhart, Hinton and Williams [10]. The BPNN learns by calculating the errors of the output layer to find the errors in the hidden layers. Due to this ability of Back-Propagating, it is highly suitable for problems in which no relationship is found between the output and inputs. Due to its flexibility and learning capabilities it has been successfully implemented in wide range of applications [11]. Although BPNN has been used successfully it has some limitations. Since it uses gradient descent learning rule which requires careful selection of parameters such as network topology, initial weights and biases, learning rate value, activation function, and value for the gain in the activation function should be selected carefully. An improper choice of these parameters can lead to slow network convergence, network error or failure. Seeing these problems, many variations in gradient descent BPNN algorithm have been proposed by previous researchers to improve the training efficiency. Some of the variations are the use of learning rate and momentum to speed-up the network convergence and avoid getting stuck at local minima. These parameters are frequently used to control the weight adjustments along the steepest descent and for controlling oscillations [12-14]. Also, Evolutionary computation is often used to train the weights and parameters of neural networks to avoid local minima. In recent years, many new techniques have been proposed for training ANN and to overcome the weakness of gradient-based techniques. These algorithms include global search techniques such as hybrid PSO-BP [15], artificial bee colony back-propagation (ABC-BP) algorithm [16-17], evolutionary artificial neural networks algorithm (EA) [18], and genetic algorithms (GA) [19] etc. But these algorithms are still not devoid of local minima problem. So, for the sake of precision and to avoid local minima in BPNN convergence, this paper proposes a new Bat-Based back-propagation (BAT-BP) algorithm which employs Bat algorithm [20] to meta-heuristically find the optimal weights in BPNN [10-14]. The proposed Bat-BP algorithm is used to train OR and XOR datasets to get null network stagnancy.

The remaining paper is organized as follows: Section 2 gives literature review on BNN. Section 3, explains Bat Algorithm. Section 4 provides the weight updating process in the proposed BAT-BP algorithm and the simulation results are discussed in section 5. And finally the paper is concluded in the Section 6.

2 Back-Propagation Neural Network (BPNN)

The BPNN has become the standard algorithm used for training multilayer perceptron. It is a generalized least mean squared (LMS) algorithm that minimizes a criterion equals to the sum of the squares of the errors between the actual and the desired outputs [10]. This principle is;

$$\mathrm{E_p} = \sum_{\mathrm{i}=1}^{\mathrm{j}} (\mathrm{e_i})^2 \tag{1}$$

where the nonlinear error signal is;

$$\mathrm{e_i} = \mathrm{d_i} - \mathrm{y_i} \tag{2}$$

d_i And y_i are, respectively, the desired and the current outputs for the ith unit. P denotes in (1) the pth pattern; j is the number of the output units. The gradient descent method is given by,

$$\mathrm{w_{ki}} = -\mu \frac{\partial \mathrm{E_p}}{\partial \mathrm{w_{ki}}} \tag{3}$$

Where w_{ki} is the weight of the ith unit in the (n-1)th layer to the kth unit in the nth layer. The BP calculates errors in the output layer ∂_l, and the hidden layer, ∂_j are using the formulas in Equation (4) and Equation (5) respectively:

$$\partial_\mathrm{l} = \mu(\mathrm{d_i} - \mathrm{y_i})\mathrm{f}'(\mathrm{y_i}) \tag{4}$$

$$\partial_\mathrm{j} = \mu \textstyle\sum_\mathrm{i} \partial_\mathrm{l} \mathrm{w_{lj}}\, \mathrm{f}'(\mathrm{y_i}) \tag{5}$$

Here d_i is the desired output of the ith output neuron, y_i is the actual output in the output layer, y_i is the actual output value in the hidden layer, and k is the adjustable variable in the activation function. The back propagation error is used to update the weights and biases in both the output and hidden layers. The weights, w_{ij} and biases, b_i, are then adjusted using the following formulae;

$$w_{ij}(k+1) = w_{ij}k + \mu\partial_j \mathrm{y_i} \tag{6}$$

$$\mathrm{b_i(k+1) = b_ik} + \mu\, \partial_\mathrm{j} \tag{7}$$

Here k is the number of the epoch and μ is the learning rate.

3 The Bat Algorithm

Bat is a metaheuristic optimization algorithm developed by Xin-She Yang in 2010[20]. Bat algorithm is based on the echolocation behavior of microbats with varying pulse rates of emission and loudness. Yang [20] has idealized the following rules to model Bat algorithm;

1) All bats use echolocation to sense distance, and they also “know” the difference between food/prey and back-ground barriers in some magical way.
2) A bat fly randomly with velocity (v_i) at position (x_i) with a fixed frequency (f_{min}), varying wavelength λ and loudness A_0 to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission $r\in$ [0,1], depending on the proximity of their target.
3) Although the loudness can vary in many ways, Yang [20] assume that the loudness varies from a large (positive) $A0$ to a minimum constant value A_{min}.

First, the initial position x_i, velocity v_i and frequency f_i are initialized for each bat b_i. For each time step t, the movement of the virtual bats is given by updating their velocity and position using Equations 8, 9 and 10, as follows:

$$f_i = f_{min} + (f_{max} + f_{min})\beta \tag{8}$$

$$v_i^t = v_i^{t-1} + (x_i^t + x_*)f_i \tag{9}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{10}$$

Where β denotes a randomly generated number within the interval [0,1]. Recall that x_i^tdenotes the value of decision variable j for bat i at time step t. The result of f_i in Equation 8 is used to control the pace and range of the movement of the bats. The variable x_* represents the current global best location (solution) which is located after comparing all the solutions among all the n bats. In order to improve the variability of the possible solutions, Yang [12] has employed random walks. Primarily, one solution is selected among the current best solutions for local search and then the random walk is applied in order to generate a new solution for each bat;

$$x_{new} = x_{old} + \in A^t \tag{11}$$

Where, A^t stands for the average loudness of all the bats at time t, and $\epsilon \in [-1,1]$ is a random number. For each iteration of the algorithm, the loudness $A_i$and the emission pulse rate r_i are updated, as follows:

$$A_i^{t+1} = \propto A_i^t \tag{12}$$

$$r_i^{t+1} = r_i^0[1 - exp(-\gamma t)] \tag{13}$$

Where α and γ are constants. At the first step of the algorithm, the emission rate, r_i^0 and the loudness, A_i^0 are often randomly chosen. Generally, $A_i^0 \epsilon [1,2]$ and $r_i^0 \epsilon [0,1]$[12].

4 The Proposed BAT-BP Algorithm

BAT is a population based optimization algorithm, and like other meta-heuristic algorithms, it starts with a random initial population. In Bat algorithm, each virtual bat flies randomly with a velocity vi at some position x_i, with a varying frequency f_i and loudness A_i, as explained in the Section IV. As, it searches and finds its prey, it changes frequency, loudness and pulse emission rate ri. Search is intensified by a local random walk. Selection of the best continues until stopping criterion are met. To control the dynamic behavior of a swarm of bats, Bat algorithm uses a frequency-tuning technique and the searching and usage is controlled by changing the algorithm-dependent parameters [20].

In the proposed BAT-BP algorithm, each position represents a possible solution (i.e., the weight space and the corresponding biases for BPNN optimization in this paper). The weight optimization problem and the position of a food source represent the quality of the solution. In the first epoch, the best weights and biases are initialized with BAT and then those weights are passed on to the BPNN. The weights in BPNN are calculated and compared in the reverse cycle. In the next cycle BAT will again update the weights with the best possible solution and BAT will continue searching the best weights until the last cycle/ epoch of the network is reached or either the MSE is achieved.

The pseudo code of the proposed Bat-BP algorithm is shown in the Figure 1:

```
Step 1: BAT is initializes and passes the best weights to BPNN
Step 2: Load the training data
Step 3: While MSE < Stopping Criteria
Step 4: Initialize all BAT Population
Step 5: Bat Population finds the best weight in Equation 9 and pass it on to the network in Equation 6 and Equation 7.
Step 6: Feed forward neural network runs using the weights initialized with BAT
Step 7: Calculate the backward error
Step 8: Bat keeps on calculating the best possible weight at each epoch until the network is converged.
End While
```

Fig. 1. Pseudo code of the proposed Bat-BP algorithm

5 Results and Discussions

Basically, the main focus of this paper is to improve the accuracy in network convergence. Before discussing the simulation test results, there are certain things that need be explained such as tools and technologies, network topologies, testing methodology and the classification problems used for the entire experimentation. The discussion is as follows:

5.1 Preliminary Study

The Workstation used for carrying out experimentation comes equipped with a 2.33GHz Core-i5 processor, 4-GB of RAM while the operating system used is Microsoft Windows 7. The simulations are carried-out using MATLAB 2010 software on three datasets such as 2-bit XOR, 3-Bit XOR and 4-bit OR. The following three algorithms are analyzed and simulated on the datasets:

1. Simple Back-Propagation Neural Network (BPNN) algorithm[10],
2. Artificial Bee Colony with Back-Propagation (ABC-BP) algorithm[16-17], and
3. The Proposed BAT based Back-Propagation (BAT-BP) algorithm

Three layer back-propagation neural networks is used for testing of the models, the hidden layer is kept fixed to 10-nodes while output and input layers nodes vary according to the datasets given. Log-sigmoid activation function is used as the transfer function from input layer to hidden layer and from hidden layer to the output layer.

For each problem, trial is limited to 1000 epochs. A total of 20 trials are run for each dataset. The network results are stored in the result file for each trial. CPU time, average accuracy, and Mean Square Error (MSE) are recorded for each independent trials on XOR and OR datasets.

5.2 XOR Dataset

The Exclusive-OR (XOR) dataset is based on the logical operation XOR which is a type of logical disjunction on two operands that results in a value of true if the operands opposite truth values. i.e., exactly one of the operands has a value of true.

2-Bit XOR Dataset

The first test problem is the 2 bit XOR Boolean function consisting of two binary inputs and a single binary output. In simulations, we used 2-10-1 network architecture for two bit XOR. For the Bat-BP, ABC-BP and BPNN, Table 1, shows the CPU time, number of epochs and the MSE for the 2 bit XOR test problem with 10 hidden neurons. Figure 2 shows the 'MSE performance vs. Epochs' of BAT-BP and ABC-BP algorithms for the 2-10-1 network architecture.

Table 1. CPU Time, Epochs and MSE for 2-bit XOR dataset with **2-10-1** ANN Architecture

Algorithms	ABC-BP	BPNN	BAT-BP
CPUTIME	19.47	13.74	**2.39**
EPOCHS	249.05	500	**23.25**
MSE	0.0019	0.2523	**0**
Accuracy (%)	98.08	75.00	**100**

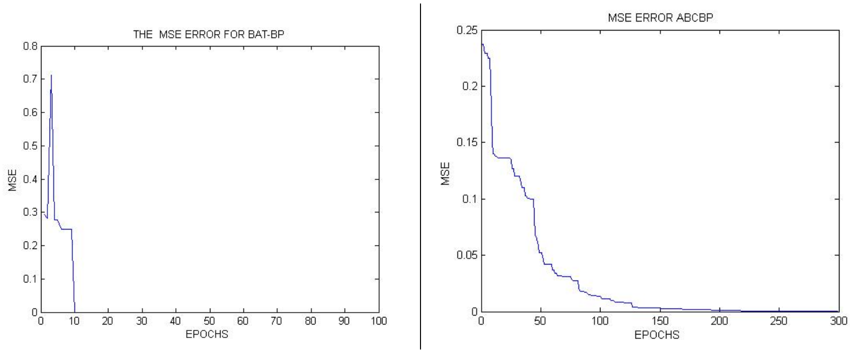


Fig. 2. (From Left to Right) Bat-BP and ABC-BP convergence performance on 2-bit XOR with 2-10-1 ANN Architecture

3-Bit XOR Dataset

In the second phase, we used 3 bit XOR dataset consisting of three inputs and a single binary output. For the three bit input we apply 3-10-1, network architecture. The parameter range is same as used for two bit XOR problem, for the 3-10-1 the network it has forty connection weights and eleven biases. For the Bat-BP, ABC-BP and BPNN, Table 2 shows the CPU time, number of epochs and the MSE for the 2 bit XOR test problem with 10 hidden neurons.

In Figure 3, we can see the simulation results 'MSE vs. Epochs' convergence performance for 3-bit XOR dataset on Bat-BP and ABC-BP algorithms. Here also, BAT-BP algorithm can be seen converging within 12 epochs, and 4.05 CPU cycles. While ABC-BP is seen converging within 21.08 CPU cycles and in 300 plus epochs. BAT-BP has slightly less accuracy and more MSE than ABC-BP this time for 3-bit XOR dataset.

Table 2. CPU Time, Epochs and MSE for 3-bit XOR dataset with **2-10-1** ANN Architecture

Algorithms	ABC-BP	BPNN	BAT-BP
CPUTIME	21.08	13.74	**4.05**
EPOCHS	283	500	**23**
MSE	0.0716	0.2523	**0.0625**
Accuracy (%)	95.937	75.00	**93.69**

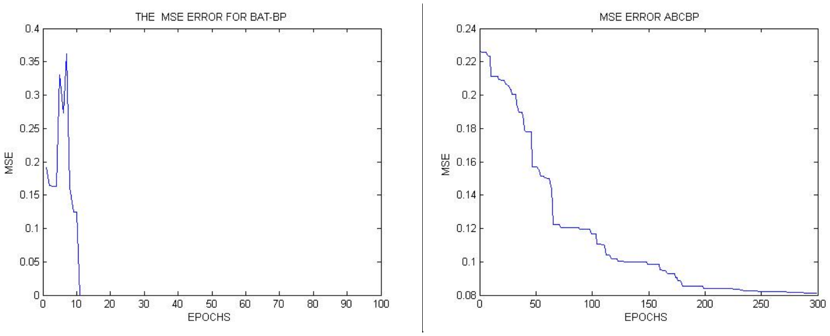


Fig. 3. (From Left to Right) Bat-BP and ABC-BP convergence performance on 3-bit XOR with 2-10-1 ANN Architecture

5.3 4-Bit OR Dataset

The third dataset is based on the logical operator OR which indicates whether either operand is true. If one of the operand has a nonzero value, the result has the value 1.

Otherwise, the result has the value 0. The network architecture used here is 4-10-1 in which the network has fifty connection weights and eleven biases. Table 3, illustrates the CPU time, epochs, and MSE performance of the proposed Bat-BP algorithm, ABC-BP, BPNN algorithms respectively. Figure 4, shows the 'MSE performance vs. Epochs' for the 4-10-1 network architecture of the proposed Bat-BP algorithm.

In Figure 5, we can see that Bat-BP is converging with a 0 MSE and 22 epochs while ABC-BP is seen converging within 42 epochs and a much higher MSE. Also, it can be noted from the Table 3 that BPNN which was failing in the previous datasets has converged to global minima with an average accuracy of 94.59 percent. For this dataset Bat-BP has again surpassed ABC-BP with an average accuracy of 100 percent.

Table 3. CPU Time, Epochs and MSE for 4-bit OR dataset with **2-10-1** ANN Architecture

Algorithms	**ABC-BP**	**BPNN**	**BAT-BP**
CPUTIME	21.17	13.32	**2.88**
EPOCHS	73.5	438	**46.8**
MSE	0.000000021	0.0546	**0**
Accuracy (%)	99.07	94.59	**100**

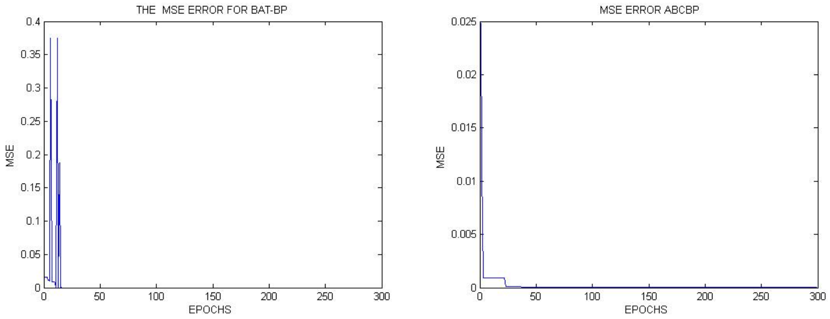


Fig. 4. (From Left to Right) Bat-BP and ABC-BP convergence performance on 4-bit OR with 2-10-1 ANN Architecture

6 Conclusions

BPNN algorithm is one of the most widely used and a popular procedure to train Artificial Neural Networks (ANN). Conventional BPNN algorithm has some drawbacks, such as getting stuck in local minima and slow speed of convergence. Nature inspired meta-heuristic algorithms provide derivative-free solution to optimize complex problems. In this paper, a new meta-heuristic search algorithm, called Bat algorithm is proposed to train BPNN to achieve fast convergence rate and accuracy. The performance of the proposed Bat-BP algorithm is compared with the ABC-BP, and BPNN algorithms. The performance of the proposed Bat-BP is verified by means of simulations on 2-bit, 3-bit XOR and 4-bit OR datasets. The simulation results show that the proposed Bat-BP converges with 0 MSE and 100 percent accuracy for 2-bit XOR and 4-bit OR datasets. Also, the CPU time is quite small as compared to ABC-BP and conventional BPNN. Further work is required to remove oscillations in the gradient descent path by introducing momentum coefficient [12] in Bat-BP algorithm. It is hoped that after introducing momentum, the CPU time, convergence rate and accuracy will become much better in Bat-BP algorithm.

References

1. Deng, W.J., Chen, W.C., Pei, W.: Back-propagation neural network based importance-performance for determining critical service attributes. J. of Expert Systems and Applications (2), 1–26 (2008)
2. Zheng, H., Meng, W., Gong, B.: Neural Network and its Application on Machine fault Diagnosis. In: ICSYSE 1992, pp. 576–579 (1992)
3. Kosko, B.: Neural Network and Fuzzy Systems, 1st edn. Prentice Hall, India (1992)
4. Basheer, I.A., Hajmeer, M.: Artificial Neural Networks: fundamentals, computing, design and application. J. of Microbiological Methods 43(1), 3–31 (2000)
5. Krasnopolsky, V.M., Chevallier, F.: Some Neural Network applications in environmental sciences. Part II: advancing computational efficiency of environmental numerical models. J. of Neural Networks 16(3), 335–348 (2003)
6. Coppin, B.: Artificial Intelligence Illuminated. Jones and Bartlet illuminated Series, USA (2004)
7. Lee, T.L.: Back-propagation neural network for the prediction of the short-term storm surge in Taichung harbor, Taiwan. J. Engineering Applications of Artificial Intelligence 21(1), 63–72 (2008)
8. Rehman, M.Z., Nawi, N.M., Ghazali, M.I.: Predicting Noise-Induced Hearing Loss (NIHL) and Hearing Deterioration Index (HDI) in Malaysian Industrial Workers using GDAM Algorithm. J. of Engineering and Technology (JET), UTeM 3(1), 179–197 (2012)
9. Nawi, N.M., Rehman, M.Z., Ghazali, M.I.: Noise-Induced Hearing Loss Prediction in Malaysian Industrial Workers using Gradient Descent with Adaptive Momentum Algorithm. International Review on Computers and Software (IRECOS) 6(5) (2011)
10. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Internal Representations by error Propagation. J. Parallel Distributed Processing: Explorations in the Microstructure of Cognition (1986)
11. Lee, K., Booth, D., Alam, P.A.: Comparison of Supervised and Unsupervised Neural Networks in Predicting Bankruptcy of Korean Firms. J. Expert Systems with Applications 29 (2005)
12. Rehman, M.Z., Nawi, N.M.: The Effect of Adaptive Momentum in Improving the Accuracy of Gradient Descent Back Propagation Algorithm on Classification Problems. CCIS Journal of Software Engineering and Computer Systems 179(6), 380–390 (2011)
13. Rehman, M.Z., Nawi, N.M., Ghazali, R.: Studying the effect of adaptive momentum in improving the accuracy of gradient descent back propagation algorithm on classification problems. International Journal of Modern Physics (IJMPCS) 1(1) (2012)

14. Nawi, N.M., Ransing, M.R., Ransing, R.S.: An improved Conjugate Gradient based learning algorithm for back propagation neural networks. J. Computational Intelligence 4 (2007)
15. Mendes, R., Cortez, P., Rocha, M., Neves, J.: Particle swarm for feed forward neural network training. In: Proceedings of the International Joint Conference on Neural Networks, vol. 2, pp. 1895–1899 (2002)
16. Nandy, S., Sarkar, P.P., Das, A.: Training a Feed-forward Neural Network with Artificial Bee Colony Based Backpropagation Method. International Journal of Computer Science & Information Technology (IJCSIT) 4(4), 33–46 (2012)
17. Karaboga, D., Akay, B., Ozturk, C.: Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 318–329. Springer, Heidelberg (2007)
18. Yao, X.: Evolutionary artificial neural networks. International Journal of Neural Systems 4(3), 203–222 (1993)
19. Montana, D.J., Davis, L.: Training feedforward neural networks using genetic algorithms. In: Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, vol. 1, pp. 762–767 (1989)
20. Yang, X.-S.: A new metaheuristic bat-inspired algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)

Probabilistic Description of Model Set Response in Neuromuscular Blockade

Conceição Rocha[1], João M. Lemos[2], Teresa F. Mendonça[1], and Maria E. Silva[3]

[1] Departamento de Matemática, Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre, 4169-007 Porto, Portugal and Center for Research &
Development in Mathematics and Applications (CIDMA),
Universidade de Aveiro, Portugal
{mnrocha,tmendo}@fc.up.pt

[2] INESC-ID/IST, Technical University of Lisbon, Lisboa, Portugal
jlml@inesc-id.pt

[3] Faculdade de Economia da Universidade do Porto, Porto, Portugal,
and Center for Research & Development in Mathematics and Applications (CIDMA),
Universidade de Aveiro, Portugal
mesilva@fe.up.pt

Abstract. This work addresses the problem of computing the time evolution of the probability density function (pdf) of the state in a nonlinear neuromuscular blockade (NMB) model, assuming that the source of uncertainty is the knowledge about one parameter. The NMB state is enlarged with the parameter, that verifies an equation given by its derivative being zero and has an initial condition described by a known pdf. By treating the resulting enlarged state-space model as a stochastic differential equation, the pdf of the state verifies a special case of the Fokker-Planck equation in which the second derivative terms vanish. This partial differential equation is solved with a numerical method based on Trotter's formula for semigroup decomposition. The method is illustrated with results for a reduced complexity NMB model. A comparison of the predicted state pdf with clinical data for real patients is provided.

Keywords: Stochastic systems, state estimation, fokker-Planck equation.

1 Introduction

The physiologic effect induced by drug administration is described by deterministic pharmacokinetic and pharmacodynamic models that represent the interaction of the drug with the patient body. These models are of compartmental type [1] and describe, for a given drug dosage, the time evolution of the plasma concentration, C_p, and the effect concentration, C_e, of the drug. Their mathematical representation consists of a system of differential equations with several unknown parameters. These dynamic processes may also be represented by reduced complexity models that, although not being compartmental modes, have

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_39, © Springer International Publishing Switzerland 2014

the advantage of leading to simpler controllers and to avoid identifiability problems because these last models have less unknown parameters [2].

Like in most practical dynamical systems, physiological effects induced by drug administration are subjects to stochastic disturbances, either internal or external. Furthermore, model parameters vary from patient to patient and, for both these reasons, anesthesia models are not deterministic. Thus, instead of computing the exact state of the system, a stochastic process that would vary from realization to realization, a probability density function (pdf) that reflects our knowledge that the state is contained in some region is to be computed. In this case, deterministic differential equations gives place to stochastic differential equations. In particular, we are interested in Markov diffusion processes modeled by stochastic differential equations and for which the pdf is a function of time that satisfies the Fokker-Planck equation (FPE) [3].

The Fokker-Planck equation is a partial differential equation (PDE) used in several fields of natural science and engineering [4–7]. In the context of Markov diffusion processes, the transition probability density of the process, *i.e.*, the time evolution of the probability density of finding the state at a given time, in a given point, is a fundamental solution of this equation.

The problem considered in this article consists of computing, as a function of time, the probability density function (pdf) of the state of a neuromuscular blockade (NMB) model given a pdf that encodes our knowledge about uncertain model parameters (that in this case depend on the patient population considered). This problem is addressed by enlarging the state with the uncertain parameter and solving a special case of the Fokker-Planck equation known as the Liouville equation [8] to propagate in time the state pdf. This PDF is solved numerically by using an algorithm that relies on Trotter's formula [9].

The contribution consists in the method to propagate the state pdf given the pdf of the uncertain parameters and its application to the NMB model. It is remarked that the method can be applied to other components of anesthesia and to other dynamic systems whose state equations depend on uncertain parameters.

The article is organized as follows. In section 2, and in order to make the text self-contained, basic notions about Markov diffucion processes, the Foker-Plank equation and Trotter's formula are reviewed. Section 3 describes the NMB model as a stochastic differential equation with uncertainty in the initial conditions corresponding to the parameter, writes the corresponding Fokker-Planck equation and presents its numeric solution. Finally, section 4 draws conclusions.

2 Diffusions and the Fokker-Planck Equation

In this section, and for the sake of clarity, the definitions as well as restrictions to the application of some of the models or equations used in the next section are presented.

2.1 Diffusion Processes

Let $\boldsymbol{X}(t)$ be a Markov process in n dimensions, described by the multidimensional stochastic differential equation (SDE) defined in the Itô sense

$$d\boldsymbol{X}(t) = \boldsymbol{f}(\boldsymbol{X}(t), t)dt + \boldsymbol{G}(\boldsymbol{X}(t), t)d\boldsymbol{W}(t), \tag{1}$$

with

$$\boldsymbol{X}(t_0) = \boldsymbol{c}, \ \ t_0 \leq t \leq T,$$

where $\boldsymbol{G}$ is $n \times d$ matrix valued function; $\boldsymbol{W}$ is an $\mathbb{R}^d$ -valued Wiener process, *i.e.*, all the coordinates $W_i(t)$ are independent one-dimensional Wiener processes; $\boldsymbol{X}$, $\boldsymbol{f}$ are n-dimensional vector valued functions and $\boldsymbol{c}$ is a random variable independent of $\boldsymbol{W}(t) - \boldsymbol{W}(t_0)$ for $t \geq 0$ [3].

Teorema 1 (Existence and Uniqueness[10]). *If the following conditions are satisfied*

1. *Coefficients are locally Lipschitz in $\boldsymbol{x}$ with a constant independent of t, that is, for every T and N, there is a constant K depending only on T and N such that for all $|\boldsymbol{x}|$,$|\boldsymbol{y}| \leq N$ and all $0 \leq t \leq T$*

$$|\boldsymbol{f}(\boldsymbol{x}, t) - \boldsymbol{f}(\boldsymbol{y}, t)| + |\boldsymbol{G}(\boldsymbol{x}, t) - \boldsymbol{G}(\boldsymbol{y}, t)| < K\,|\boldsymbol{x} - \boldsymbol{y}|\,,$$

   then for any given $\boldsymbol{X}(0)$ the strong solution to SDE is unique.
2. *The linear growth condition holds*

$$|\boldsymbol{f}(\boldsymbol{x}, t)| + |\boldsymbol{G}(\boldsymbol{x}, t)| \leq K_T(1 + |\boldsymbol{x}|),$$

   $\boldsymbol{X}(0)$ is independent of $\boldsymbol{W}$, and $E\,|\boldsymbol{X}(t_0)|^2 < \infty$,

then the strong solution exists and is unique on $[t_0, T]$.

If the conditions of the above existence and uniqueness theorem are satisfied for the SDE (1) and in addition the functions $\boldsymbol{f}$ and $\boldsymbol{G}$ are continuous with respect to t, the solution $\boldsymbol{X}(t)$ is a n-dimensional diffusion process on $[t_0, T]$ with drift vector $\boldsymbol{f}$ and diffusion matrix $\boldsymbol{b} = \boldsymbol{G}\boldsymbol{G}^T$, with $\boldsymbol{G}^T$ denoting the transposed of $\boldsymbol{G}$.

2.2 Fokker-Plank Equation

A property of diffusion processes is that their transition probability is, under certain regularity assumptions, uniquely determined merely by the drift vector and the diffusion matrix.

Teorema 2 ([3]). *Let $\boldsymbol{X}(t)$, for $t_0 \leq t \leq T$, denote a n-dimensional diffusion process with a transition density $p(s, \boldsymbol{x}, t, \boldsymbol{y})$. If the derivatives $\partial p/\partial t$, $\partial(f_i(t, \boldsymbol{y})p)/\partial y_i$ and $\partial^2(b_{ij}(t, \boldsymbol{y})p)/\partial y_i \partial y_j$ exist and are continuous functions, then, for fixed s and $\boldsymbol{x}$ such that $s \leq t$, this transition density is a fundamental solution of the Fokker-Planck equation*

$$\frac{\partial p}{\partial t} + \sum_{i=1}^{n} \frac{\partial(f_i(t, \boldsymbol{y})p)}{\partial y_i} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2(b_{ij}(t, \boldsymbol{y})p)}{\partial y_i \partial y_j} = 0. \tag{2}$$

The boundary condition for Eq.(2) is given by $\lim_{t \to s} p(s, \boldsymbol{x}, t, \boldsymbol{y}) = \delta(\boldsymbol{y} - \boldsymbol{x})$.

This partial differential equation (PDE) has an analytical solution only in some special cases and, in general, numerical methods are need to solve it. In this work a method based on Trotter's formula for semigroup decomposition, explained below, is used.

2.3 Semigroup Definition

Consider the Banach space $\boldsymbol{X}$ of continuous functions equipped with the supremum norm.

Definition 1 ([11]). *A semigroup of operators of class C_0 is a family of operators $T_t$defined in $\boldsymbol{X}$ and indexed by the parameter $t \in \mathbb{R}$ (time) such that:*

1. *T_t is defined* $\forall t \geq 0$*;*
2. *T_t satisfies the* semigroup condition*:*

$$\forall_{s,t \in \mathbb{R}}\, T_{t+s} = T_t T_s \tag{3}$$

3. *T_t satisfies the continuity condition*

$$\lim_{t \to \infty} T_t \boldsymbol{x} = \boldsymbol{x}\ \forall_{\boldsymbol{x} \in \boldsymbol{X}}$$

4. *T_t is bounded* $\forall t \geq 0$:

$$\exists_{c \in \mathbb{R}} : \forall_{\boldsymbol{x} \in \boldsymbol{X}} \|T\boldsymbol{x}\| \leq c \|\boldsymbol{x}\|$$

Definition 2 ([11]). *The* infinitesimal generator *of the semigroup T_t is the operator defined by*

$$A = \lim_{t \to 0} t^{-1}(T_t - I)$$

where I is the identity operator.

Remark 1. *The set $B(\boldsymbol{X})$ of bounded linear operators in a Banach space $\boldsymbol{X}$ is itself a Banach space with respect to the norm induced by the norm defined in $\boldsymbol{X}$:*

$$\|T_t\| \triangleq sup \left\{ \frac{\|T_t \boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \boldsymbol{x} \in \boldsymbol{X}\ \ \{0\} \right\}$$

Under this norm, definition 2 states that the semigroup T_t satisfies the following so called evolution equation

$$\frac{d}{dt}T_t = AT_t \tag{4}$$

with the initial condition $T_0 = I$. The solution T_t of (4) is referred to as the integral operator corresponding to a A.

2.4 Trotter's Formula

Consider the situation in which A is the sum of two operators A_1 and A_2. Let T_t^1 and T_t^2 be the corresponding integral operators (semigroups), *i. e.*, assume that

$$\frac{d}{dt}T_t^i = A_i T_t^i,\ i = 1, 2 \tag{5}$$

with T_t satisfying the evolution equation

$$\frac{d}{dt}T_t = (A_1 + A_2)T_t. \tag{6}$$

In general, it is not true that T_t results from the composition of T_t^1 and T_t^2. However, this is approximately true for small t, meaning that T_t can be approximated by the iterated composition of T_Δ^1 ans T_Δ^2 over small intervals of time Δ. This is stated in the following theorem:

Teorema 3 ([9]). *Let T_t^1 and T_t^2 satisfy the* norm condition*:*

$$\exists_{w \in \mathbb{R}} : \forall_{t>0} \left\| T_t^i \right\| \leq e^{w_i t},\ i = 1, 2$$

and that $D(A_1 + A_2) = D(A_1) \cap D(A_2)$ is dense in $\boldsymbol{X}$, where $D(A)$ denotes the domain of A. Then, (the closure of) $A_1 + A_2$ generates a semigroup of class C_0 iff (the closure) $R(\lambda I - A_1 - A_2)$ is dense in $\boldsymbol{X}$ for some $\lambda > w_1 + w_2$, where $R(A)$ denotes the range of A. If $A_1 + A_2$ (or its closure) generates a semigroup of class C_0, this is given by

$$T_t = \lim_{\Delta \to 0} (T_\Delta^1 T_\Delta^2)^{\lceil t/\Delta \rceil} \tag{7}$$

where $\lceil t/\Delta \rceil$ represents the greatest integer that does not exceed t/Δ.

Expression (7) is commonly known as Trotter's formula. It embodies an approximation that may be extended to a finite sum of operators.

3 Transition Probability in NMB

The neuromuscular blockade dynamics can be represented by a Wiener model comprising a linear state-space model and a nonlinear output equation. The influence of the parameter uncertainty on NMB state and output (the NMB level) is studied hereafter using the method previously described.

3.1 NMB Dynamics

Recently a reduced complexity model for the neuromuscular blockade induced by *Atracurium* was proposed [2] that has compartmental features and is represented by

$$\begin{cases} \dot{x}_1 = -k_3\alpha x_1 \\ \dot{x}_2 = \ \ k_2\alpha x_1 \ \ -k_2\alpha x_2 \\ \dot{x}_3 = \qquad\qquad\quad k_1\alpha x_2 \ \ -k_1\alpha x_3 \end{cases} \tag{8}$$

Here the dot denotes the time derivative; k_1, k_2 and k_3 are known process parameters; x_1, x_2 and x_3 are state variables; and α is an unknown model parameter. The advantage of this model consists in the fact that the description of inter-patient variability is reduced to the unknown parameter α, considered to be a random variable described by a probability density function. Therefore, all state variables are random outputs and the system can be rewritten as a stochastic system with a state enlarged by the parameter, as

$$\begin{cases} \dot{x}_1 = -k_3\alpha x_1 \\ \dot{x}_2 = \ \ k_2\alpha x_1 \ \ -k_2\alpha x_2 \\ \dot{x}_3 = \qquad\qquad\quad k_1\alpha x_2 \ \ -k_1\alpha x_3 \\ \dot{\alpha} = \qquad 0 \end{cases} \tag{9}$$

or

$$d\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}(t), \alpha(t), t)dt \tag{10}$$

with $\boldsymbol{f}$ defined from (9), and

$$d\alpha = 0dt \tag{11}$$

with initial conditions $\boldsymbol{X}_0 = [x_1(t_0), x_2(t_0), x_3(t_0)]^T$ and $\alpha = \alpha(t_0)$ a random variable with a known pdf.

Since the conditions of theorem 1 (Existence and Uniqueness) are verified and the functions f_i are continuous, an equivalent description can be given in terms of a three-dimensional Fokker-Planck equation, and the time propagation of the probability density function of state variables obtained

$$\tfrac{\partial p}{\partial t} = k_3\alpha(p + x_1\tfrac{\partial p}{\partial x_1}) + k_2\alpha(p - (x_1 - x_2)\tfrac{\partial p}{\partial x_2}) + k_1\alpha(p - (x_2 - x_3)\tfrac{\partial p}{\partial x_3}) \tag{12}$$

with boundary condition $\lim_{t\to 0} p(x_1, x_2, x_3, \alpha, t) = p(x_1, x_2, x_3, \alpha, 0)$.

Actually, (12) is a degenerate form of the Fokker-Planck equation (Liouville Equation [8]) because the second derivative term associated to diffusion is assumed to vanish. The solution of (12) represents how the state pdf is influenced by the pdf of the parameter α and evolves along time. A numerical method based on Trotter's formula is applied hereafter in order to approximate the solution of (12). For that purpose, (12) is rewritten as

$$\frac{\partial p(x_1, x_2, x_3, \alpha, t)}{\partial t} = (L_1 + L_2 + L_3 + L_4)p(x_1, x_2, x_3, \alpha, t) \tag{13}$$

where the infinitesimal generators L_1, L_2, L_3 and L_4 are defined by

$$L_1 p(x_1, x_2, x_3, \alpha, t) = (k_1 + k_2 + k_3)\alpha p(x_1, x_2, x_3, \alpha, t) \tag{14}$$

$$L_2 p(x_1, x_2, x_3, \alpha, t) = k_3 \alpha x_1 \frac{\partial p(x_1, x_2, x_3, \alpha, t)}{\partial x_1} \tag{15}$$

$$L_3 p(x_1, x_2, x_3, \alpha, t) = k_2 \alpha (x_2 - x_1) \frac{\partial p(x_1, x_2, x_3, \alpha, t)}{\partial x_2} \tag{16}$$

$$L_4 p(x_1, x_2, x_3, \alpha, t) = k_1 \alpha (x_3 - x_2) \frac{\partial p(x_1, x_2, x_3, \alpha, t)}{\partial x_3} \tag{17}$$

The operators T_t^i generated by the infinitesimal generators L_i, $i = 1, 2, 3, 4$ are given by

$$T_\Delta^1 p(x_1, x_2, x_3, \alpha, t) = e^{\alpha(k_1+k_2+k_3)\Delta} p(x_1, x_2, x_3, \alpha, t) \tag{18}$$

$$T_\Delta^2 p(x_1, x_2, x_3, \alpha, t) = p(x_1 e^{-k_3 \alpha \Delta}, x_2, x_3, \alpha, t) \tag{19}$$

$$T_\Delta^3 p(x_1, x_2, x_3, \alpha, t) = p(x_1, x_1 + e^{-k_2 \alpha \Delta}(x_2 - x_1), x_3, \alpha, t) \tag{20}$$

$$T_\Delta^2 p(x_1, x_2, x_3, \alpha, t) = p(x_1, x_2, x_2 + e^{-k_1 \alpha \Delta}(x_3 - x_2), \alpha, t) \tag{21}$$

Since all the operators satisfy the conditions of definition 1 as well as the norm condition of theorem 3 is valid to apply Trotter's formula. Accordingly, the solution of (12) is approximated by

$$p(x_1, x_2, x_3, \alpha, t + \Delta) \approx T_\Delta^1 T_\Delta^2 T_\Delta^3 T_\Delta^4 p(x_1, x_2, x_3, \alpha, t), \tag{22}$$

meaning that

$$\begin{aligned} p(\boldsymbol{x}, \alpha, t + \Delta) \approx e^{\alpha(k_1+k_2+k_3)\Delta} & p(x_1 e^{-k_3 \alpha \Delta}, \\ & x_1 + e^{-k_2 \alpha \Delta}(x_2 - x_1), x_2 + e^{-k1 \alpha \Delta}(x_3 - x_2), \alpha, t). \end{aligned} \tag{23}$$

3.2 State Uncertainty Characterization

In order to illustrate the results, start by addressing a simplified one-dimensional case. Two parameter distributions are considered, namely the *lognormal* (LN) and the *uniform* distribution (U) defined as:

- For the *lognormal* distribution

$$f(\alpha) = \frac{1}{\sqrt{2\pi}\sigma\alpha} exp\left\{ -\frac{(ln(\alpha) - \mu)^2}{2\sigma^2} \right\}$$

  with $\mu = -3.287$ and $\sigma = 0.158$.
- For the *uniform* distribution

$$f(\alpha) = \begin{cases} 1/(b-a) \ , & for \quad a \le \alpha \le b \\ \quad 0 \quad\ , & for \ \alpha < a \ or \ \alpha > b \end{cases},$$

  with $a = 0.027$ and $b = 0.052$.

The four parameters used in the two distributions are the maximum likelihood estimates for a real database of patient data with 48 samples. To apply Trotter's formula the interval $\varDelta$ is made constant and equal to 0.1 minute.

One Dimensional Case. Before computing the impact of the state uncertainty on the system output (measured NMB level), and for the sake of illustration in a simple case, consider the one-dimensional case, in which

$$k_1 = 0,\ k_2 = 0 \text{ and } k_3 = 10$$

and the initial condition is $x_1(0) = 500k_3\alpha$ with initial probability density function

$$p(x_1, \alpha, 0) = f_\alpha(\alpha)\delta(x_1 - x_1(0))$$

where $f_\alpha(\alpha)$ is the probability density function of the parameter α.

First, the time evolution of the probability density function induced by each one of the two operators used in the Fokker-Planck equation is computed separately. Then, the approximated solution yielded by Trotter's formula, *i.e.*, the time evolution of the probability density function induced by the two operators, is represented and discussed.

The operator L_1 acts in the transition probability by means of one factor that depends on the value of α. This action deforms the transition probability by increasing pointwise the pdf, but does not change the position of the pdf to which it is applied, with respect to the values of x_1. Instead, the operator L_2 acts in the transition probability by causing a shift and a change of the independent variable *i.e.*, this operator replaces x_1 by $x_1 e^{-k_3\alpha t}$. When the two operators are applied in sequence, the result is represented on figure 1.

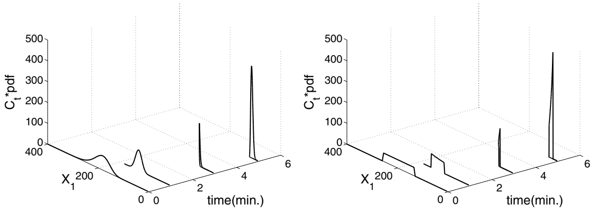


Fig. 1. Action of both operators, $L_1 + L_2$-parameter distribution LogN (left) and uniform distribution (right)

Neuromuscular Blockade. The NMB level is computed from the state variable using the output equation. This equation is nothing more than a static function that allows to compute the NMB level r as function of one of the state variables [2]. Therefore, the NMB signal pdf as a function of time t is computed using a pdf transformation associated to the output function.

Figure 2 shows the NMB pdf at 6 different time instants. In the plane [r, t] a set of responses from 13 real patients (clinical results) are also plotted.

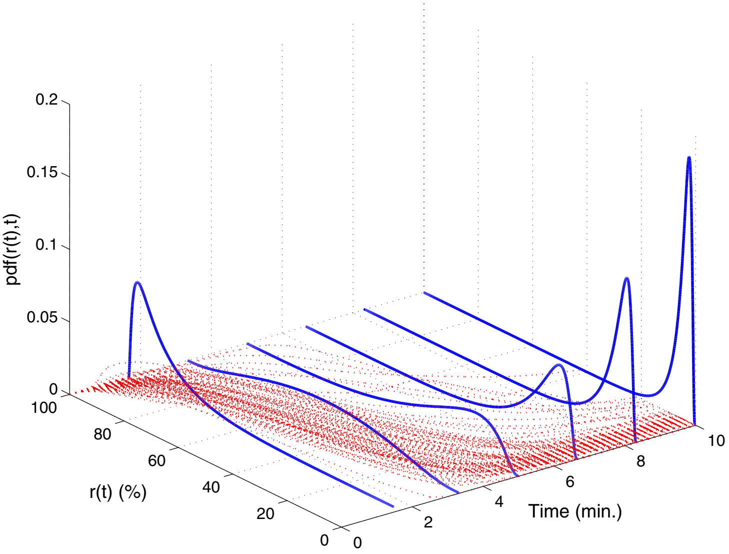


Fig. 2. Neuromuscular blockade pdf as computed from the model and a set of 13 responses from real patients

4 Conclusions

This work allows to see that the physiological effect induced by *atracurium* administration has different density transition probability for different parameter distribution. Moreover, the range of values for the state variables that may occur depends not only of the parameter distribution but also on the instance under consideration.

In this problem, Trotter's formula provides an adequate approximation for the transition probability given by the solution of the Fokker-Planck equation for this stochastic system.

The time evolution of the transition density probability to the administration of an *atracurium* bolus of 500 $\mu g/kg$ (that corresponds to the usual procedure at the beginning of a general anesthesia), given by the solution of the Fokker-Planck equation is in accordance with the expected. This means that, for the same drug dosage applied, different patient have states that evolve in time in a different way. Nevertheless, all the states will converge for zero, and that is also expected since the drug will be eliminated from the body of the patient.

This work shows that the parameters uncertainty has an important role in the states uncertainty, and it is immediately after the drug administration that it is most noted. For further work the authors intend to study the influence of the parameters uncertainty over time, assuming that the unknown parameter instead of being constant in time is affected by disturbances. This may be seen as a stochastic approach to the on-line parameter identification problem.

Acknowledgment. This work was supported by *FEDER* funds through *COMPETE*–Operational Programme Factors of Competitiveness and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (University of Aveiro) and "FCT–Fundação para a Ciência e a Tecnologia", within projects PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690, PEst-OE/EEI/LA0021-/2013 and GALENO - Modeling and Control for personalized drug administration, PTDC/SAU-BEB/103667/2008. Conceição Rocha acknowledges the grant SFRH/BD/61781 /2009 by FCT/ESF.

References

1. Bailey, J., Haddad, W.: Drug dosing in clinical pharmacology: paradigms, benefits, and challenges. IEEE Control Syst. Mag. 25(2), 35–51 (2005)
2. Silva, M.M., Wigren, T., Mendonça, T.: Nonlinear identification of a minimal neuromuscular blockade model in anesthesia. IEEE Transactions on Control Systems Technology 20(1), 181–188 (2012)
3. Arnold, L.: Stochastic Differential Equations: Theory and Applications. John Wiley & Sons, New York (1974)
4. Fall, C., Marland, E., Wagner, J., Tyson, J. (eds.): Computational Biology. Springer, New York (2002)
5. Grindrod, P.: The Theory and Applications of Reaction-Diffusion Equations – Patterns and Waves. Clarendon Press, Oxford (1996)
6. Mukherjee, A., Strikwerda, J.C.: Analysis of dynamic congestion control protocols – a fokker-planck approximation. In: Proc. ACM Sigcomm, Zurich, Switzerland, pp. 159–169 (1991)
7. Lemos, J.M., Moura, J.M.F.: Time sampling of diffusion systems using semigroup decomposition methods. In: MTNS 2004, 16th Int. Symp. on Mathematical Theory of Networks and Systems, Leuven, Belgium (2004)
8. Brockett, R.: Notes on the Control of the Liouville Equation. In: Cannarsa, P., Coron, J.-M. (eds.) Control of Partial Differential Equations, ch. 2. Springer (2010)
9. Trotter, H.F.: On the product of semigroups of operators. Proc. American Mathematical Society 10(4), 545–551 (1959)
10. Klebaner, F.C.: Introdution to Stochastic Calculus With Applications. Imperial College Press, London (2005)
11. McBride, A.C.: Semigroups of linear operators: An introdution. Longman Scientific & Technical, London (1987)

Retracted: Matlab Simulation of Photo Propagation in Three-Layer Tissue

Julia Kurnatova[1], Dominika Jurovata[1], Pavel Vazan[1], and Peter Husar[2]

[1] Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava, Hajdóczyho 1, Trnava 917 24, Slovak Republic
[2] Ilmenau University of Technology, Institute of Biomedical Engineering and Informatics, 98693 Ilmenau, Germany

Abstract. This paper deals with the simulation of photon propagation in the maternal abdomen. Authors focused on the light transport, photon trajectory and their radiation in three-layer tissue. The main aim of this study is to observe the behaviour of photon in three-layer tissue. A simulation model has been implemented in Matlab. The photon interaction with tissue was observed. This model was realized for the project aimed to non-invasive pulse oximetry measurement of fetal oxygen saturation in the maternal abdomen. One of the fundamental challenges is to ensure a sufficient penetration depth which covers maternal and fetal tissue. This contribution investigates the photon trajectories and compares the results of specular reflectance, diffuse reflectance, absorbed fraction and transmittance in three-layer tissue with regard to the thickness of the third layer. Simulations have been performed at three depths fetal (2.5, 3.7, 4.9 cm).

1 Introduction

Optical technique is a useful analysis method in biomedical diagnostics and monitoring of biological tissues such as brain imaging and for fetal heart rate detection and oxygen saturation measurement due to its theoretical advantages in comparison with other modalities. Interaction of laser light with turbid medium (e.g. human tissue) depends on the optical properties of the medium i.e. refractive index n, absorption coefficient μ_a, scattering coefficient μ_s and anisotropy factor g. The simulation of light transport provides statements about the photon interaction with tissue. In the case of fetal pulse oximetry, it will be possible to evaluate the light distribution and the penetration depth under different conditions without the need of suitable patients. These parameters are important for further investigations, for instance simulating pulse curve shapes or determining the oxygenation of arterial blood. The behaviour of the photon migration process in turbid media is a fundamental research in many practical applications in

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_40, © Springer International Publishing Switzerland 2014

biological tissue. For these reasons this paper deals with the fundamental photon propagation rules and the radiation in tissue.

2 The Photon Propagation Rules

The implemented algorithm is based on Wang and Jacques steady-state light transport model which was written a in a standard language C [7]. We used the functions of propagation rules.

Figure 1 shows the schematic of the Cartesian coordinate system which describes the model. The z-coordinate represents the depth of the tissue, where the x and y direction are assumed as infinity wide.

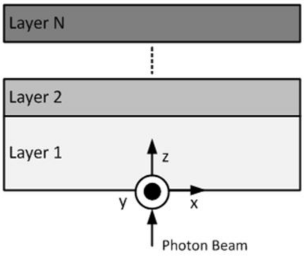


Fig. 1. Schematic of the Cartesian coordinate system, which describes the implemented simulation model, the y axis points outwards [7]

Launching a Photon

The start position of each photon is determined by the coordinates (x,y,z)=(0,0,0) and the initial direction is orthogonal to the tissue surface, which is given by $(\mu_x, \mu_y, \mu_z,)$. When the photons penetrate into the tissue, some specular reflectance at the surface will occur. The specular reflectance R_{sp} can be described by:

$$R_{sp} = \frac{(n_1 - n_2)^2}{(n_1 + n_2)^2}$$

and the photon weight will be decreased by R_{sp}

$$W = 1 - R_{sp}$$

Moving the Photon

After photon injection the step size s will be calculated by using the equation:

$$s = \frac{-ln\left(\xi\right)}{\mu_t}$$

where μ_t is an interaction coefficient equals the sum of the absorption coefficient μ_a and scatter coefficient μ_s. The parameter ξ is a random variable, which is uniformly distributed over the interval $(0, 1)$. A decision has to be made, which distinguishes whether the step size s is long enough to reach a boundary or not. If the photon did not reach a boundary the position of the photon will be updated by:

$$x \leftarrow x + \mu_x \cdot s$$
$$y \leftarrow y + \mu_y \cdot s$$
$$z \leftarrow z + \mu_z \cdot s$$

Absorption and Scattering of the Photon

By moving a photon inside the tissue the photon weight is decreasing due to absorption. The amount of photon weight loss is defined by:

$$\triangle W = W \cdot \frac{\mu_a}{\mu t}$$

The photon weight is then updated by:

$$W \leftarrow W - \triangle W$$

For scattering the photon, the azimuth $\psi \in [0, 2\pi)$ and deflection angle $\theta \in [0, \pi)$ have to be taken into account. The final photon directions are computed by the following equation:

$$\mu'_x = \frac{\sin\theta}{\sqrt{1-\mu_z^2}}\left(\mu_x\mu_z\cos\psi - \mu_y\sin\psi\right) + \mu_x\cos\theta$$
$$\mu'_y = \frac{\sin\theta}{\sqrt{1-\mu_z^2}}\left(\mu_y\mu_z\cos\psi - \mu_x\sin\psi\right) + \mu_y\cos\theta$$
$$\mu'_z = -\sin\theta\cos\psi\sqrt{1-\mu_z^2} + \mu_z\cos\theta$$

For the special case, that the incident angle is orthogonal to the surface of the tissue, the photon direction is following the formulas:

$$\mu'_x = \sin\theta\cos\psi$$
$$\mu'_y = \sin\theta\cos\psi$$
$$\mu'_z = SING\left(\mu_y\right)\cos\theta$$

$$SIGN\left(x\right) = \begin{cases} -1 \text{ if } x < 0 \\ \ \ 0 \text{ if } x = 0 \\ \ \ 1 \text{ if } x > 0 \end{cases}$$

Finally the photon direction is updated:

$$\mu_x \leftarrow \mu'_x$$
$$\mu_y \leftarrow \mu'_y$$
$$\mu_z \leftarrow \mu'_z$$

Reflection and Transmission at a Boundary

If the step size is long enough to hit the boundary, then the photon moves to the boundary. Subsequently the program decides whether the photon escapes the tissue or is internally reflected. This depends on the angle of incidence α_i and the angle of transmission α_t. The internal reflectance $R\left(\alpha_i\right)$ is then calculated by Fresnels formula:

$$R(\alpha_i) = \frac{1}{2}\left[\frac{\sin^2(a_i - a_t)}{\sin^2(a_i + a_t)} + \frac{\tan^2(a_i - a_t)}{\tan^2(a_i + a_t)}\right].$$

The finally decision is realized by comparing the internal reflectance with a random number. After this step the absorption and scattering will computed correspondingly (see [7] for more details).

Photon Termination

A photon is terminated if it escapes the tissue or if the photon weight decreases below a defined threshold inside of the tissue. In the case that the photon weight is lower than the threshold, the current photon gets a further chance in m (e.g., $m = 10$) for surviving with a weight of mW [7]. The photon is terminated if it does not survive the so called roulette:

$$W = \begin{cases} mW \text{ if } \xi \leq 1 \setminus m \\ 0 \text{ if } \xi > 1 \setminus m \end{cases}$$

3 Three-Layered Tissue Model

The anatomical model (Fig. 2) shows the mother's abdominal tissue, amniotic fluid and fetus. Previous studies have outlined the use of the perturbation method in model photon transport through an 8 cm diameter fetal brain located at a constant 2.5 cm below a curved maternal abdominal surface with an air/tissue boundary [2]. In order to study the photon-migration process, the anatomical model has been simplified into a three-layered tissue model which has been reported in the literature [6] [9] [8]. A three-layered tissue model consists of maternal d_M, amniotic fluid d_{am} and fetal layers d_F. This model is represented in Fig. 2. Maternal layer thickness and amniotic fluid layer thickness in this model are obtained from the literature [1].

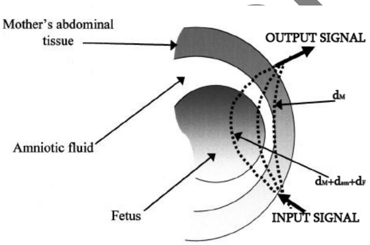


Fig. 2. A three-layered tissue model for light transport

In this work, the fetal layer thickness is given like infinite thickness. It is close to the real-life conditions, because it is not known in advance how will be the fetus turned. As a result, light impinged on the fetal tissue will penetrate and travel into an unknown depth. Therefore, a finite fetal layer thickness (given in [6]) is not an appropriate boundary condition to perform the simulation.

Table 1 shows the optical properties (absorption, scattering and refraction index) of the tissue model which are obtained from the previous study. The aim of this simulation is to estimate the number of photons remaining in the tissue and the number of photons that leave the tissue a three-layered tissue model.

Simulations have been performed at 2.5, 3.7 and 4.9 cm fetal depths. Fetal depth is defined as the total thickness of the maternal and the amniotic fluid layer. Fifty thousand photons were selected to run the simulation.

Table 1. Optical property of the proposed tissue model

	Description	**Symbol**	**Values**	**Units**	**References**
Mother layer (M)	Absorption coefficient	$\mu_a(M)$	0.08	cm^{-1}	[6]
	Reduced scattering	μ_s'	5	cm^{-1}	[6]
	Anisotropy	g	0.8	NA	[4]
	Refractive indices	n_s	1.3	NA	[3]
	Average maternal layers path length	d_M	2.4 ± 0.8	cm	[5]
Amniotic fluid layer (am)	Absorption coefficient	$\mu_a(am)$	0.02	cm^{-1}	[6]
	Reduced scattering	μ_s'	0.1	cm^{-1}	[6]
	Anisotropy	g	0.85	NA	[6]
	Refractive indices	n_s	1.3	NA	[3]
	Average maternal layers path length	d_{am}	1.3 ± 0.4	cm	[5]
Fetal layer (F)	Absorption coefficient	$\mu_a(F)$	0.125	cm^{-1}	[6]
	Reduced scattering	μ_s'	5	cm^{-1}	[6]
	Anisotropy	g	0.8	NA	[4]
	Refractive indices	n_s	1.3	NA	[3]
	Average maternal layers path length	d_F	∞	cm	[6]

4 Simulation Results

The first simulation was performed in depth 2.5 cm fetal layer. The results of 50 thousand photons were as follows Table 2.

Table 2. The results in depth 2.5 cm fetal layer

Specular reflectance	0.0170132
Diffuse reflectance	0.31962
Absorbed fraction	0.422297
Transmittance	0.24165

Number of Photons that finished out of tissue was 27718.

Number of Photons that stayed in tissue was 22282.

The second simulation was performed in depth 3.7 cm fetal layer. The results of 50 thousand photons are shown in Table 3.

Table 3. The results in depth 3.7 cm fetal layer

Specular reflectance	0.0170132
Diffuse reflectance	0.305283
Absorbed fraction	0.449149
Transmittance	0.229135

Number of Photons that finished out of tissue was 27139.

Number of Photons that stayed in tissue was 22861.

The last simulation was performed in depth 4.9 cm fetal layer. The results of 50 thousand photons were as follows Table 4.

Table 4. The results in depth 4.9 cm fetal layer

Specular reflectance	0.0170132
Diffuse reflectance	0.302399
Absorbed fraction	0.478362
Transmittance	0.202865

Number of Photons that finished out of tissue was 27624.

Number of Photons that stayed in tissue was 22376.

Fig. 3 shows the photon interaction with the tissue for a single photon. These results are based on the three-layered tissue simulation where the corresponding parameters are given by Table 1. The starting point of the photon is indicated with a green point at coordinates (0; 0). The ending point represent red point.

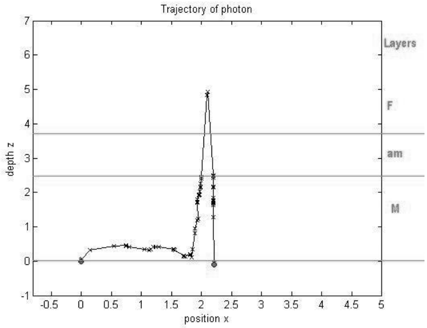


Fig. 3. Trajectory of a single photon in three-layer tissue

5 Conclusion

Based on the data obtained from the experiments, we can conclude that the specular reflectance can be generalized to any thickness of tissue to 0.0170132. Absorbed fraction rises by an increasing thickness of the third layer. Diffuse transmittance and reflectance are on the decrease with increasing the thickness of fetus.

References

[1] Gan, K.B., Zahedi, E., Mohand Ali, M.A.: Investigation of optical detection strategies for transabdominal fetal heart rate detection using three-layered tissue model and Monte Carlo simulation. Universiti Kebangsaan Malaysia, Bangi (2011)

[2] Jacques, S.L., Ramanujam, N., Vishnoi, G., Choe, R., Chance, B.: Modeling photon transport in transabdominal fetal oximetry. Journal of Biomedical Optics 5(3), 277–282 (2000)

[3] Mannheimer, P.D., Casciani, J.R., Fein, M.E., Nierlich, S.L.: Wavelength selection for low-saturation pulse oximetry. IEEE Transactions on Biomedical Engineering 44(3), 148–158 (1997)

[4] Reuss, J.L.: Arterial pulsatility and the modeling of reflectance pulse oximetry. In: Proceedings of the 25th Annual International Conference of the IEEE EMBS, Cancun, Mexico, pp. 2791–2794 (2003)

[5] Richards, D.S., Allen, S.G., White, M.A., Perez, D.R.: Umbilical vessels: Visualization Department of Obstetrics and Gynecology. University of Florida College of Medicine (1992)

[6] Ramanujam, N., Vishnoi, G., Hielscher, A.H., Rode, M.E., Forouzan, I., Chance, B.: Photon migration through the fetal head in utero using continuous wave, near infrared spectroscopy: Clinical and experimental model studies. Journal of Biomedical Optics 5(2), 173 (2000)

[7] Wang, L., Jacques, S.L.: Monte Carlo Modeling of Light Transport in Multi-layered Tissues in Standard C. University of Texas, Houston (1992)

[8] Zahedi, E., Beng, G.K.: Applicability of adaptive noise cancellation to fetal heart rate detection using photoplesthysmography. Computers in Biology and Medicine 38(1), 31–41 (2008)

[9] Zourabian, A., Siegel, A., Chance, B., Ramanujam, N., Rode, M., Boas, D.A.: Trans-abdominal monitoring of fetal arterial blood oxygenation using pulse oximetry. Journal of Biomedical Optics 5(4), 391–405 (2000)

Magnetorheological Damper Dedicated Modelling of Force-Velocity Hysteresis Using All-Pass Delay Filters

Piotr Krauze and Janusz Wyrwał

Institute of Automatic Control, Silesian University of Technology
{piotr.krauze,janusz.wyrwal}@polsl.pl

Abstract. The paper presents a novel approach to the problem of force-velocity characteristics modelling dedicated to MR dampers. It is stated that velocity and control dedicated dynamic signal paths need to be included in MR damper model. It is shown that hysteretic behaviour may be modelled using all-pass delay filters located in the velocity dedicated signal path. Parameters of the presented model are estimated using measurement data obtained by means of Material Testing System (MTS). Experiments are performed for damper excitation frequencies assumed within range of 0.5 Hz – 2.5 Hz and control current levels restricted within 0.05 A – 1.0 A. Parameters of delay filters are estimated and accuracy of the reference acceleration based hysteresis model and referred model based on delay filters are compared. Results demonstrate that delay filters based model maps MR damper dynamics, mainly hysteretic behaviour, with high accuracy.

Keywords: Magnetorheological damper, behavioural model, hysteretic behaviour, all-pass delay filters.

1 Introduction

Intelligent materials, such as magnetorheological (MR) fluids, play an important role in real-time control of mechanical structures and devices which are burdened with vibrations. Starting from large structures such as cable-stayed bridges [1] and high buildings [2], MR dampers are used to suppress vibrations which are induced by wind and rain as well as earthquakes. In case of medium-sized road vehicles with semiactive shock-absorbers located in seats and/or vehicle suspension systems [3, 4] the ride comfort and ride safety can be significantly improved by adaptive change of damper dynamic viscosity compared with passive solutions. Other authors [5] present an application of smart disc damper applied in rotor system used for vibration control. MR dampers seem to be also competitive compared to mechanically adjustable dampers due to quicker dynamic response.

MR dampers based vibration control systems which are installed in vehicles are favoured considering low energy consumption and inherent stability. On the other hand, semiactivity of MR dampers makes it impossible to introduce energy into the suspension system they are applied to. Consequently, semiactive

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_41, © Springer International Publishing Switzerland 2014

systems can dissipate vibration energy only. Constraints of semiactive dampers require complex control schemes [3] to be used in vibration control. Moreover, strongly nonlinear relationship between relative piston velocity and force makes MR dampers challenging elements to control. Numerous phenomena such as pre-yield and post-yield regions presented in [4, 5], biviscous characteristics [4] and dynamics of the velocity to force and control signal to force signal paths, which result in hysteretic behavior [4, 6, 7], need to be included in MR damper models.

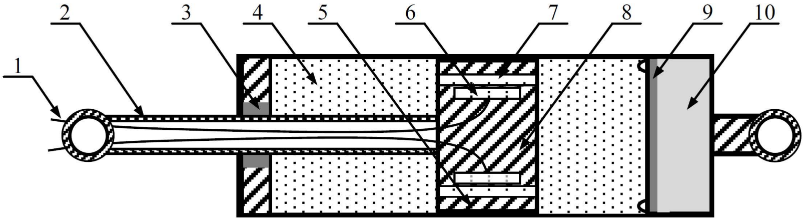


Fig. 1. MR damper structure: 1- coil wires, 2- piston rod, 3- bearing and seal, 4- MR fluid, 5- ring, 6- coil, 7- orifice, 8- piston, 9- diaphragm, 10- gas accumulator

MR dampers consist of coupled elements, i.e., cylidrical housing, rod and a specially designed piston (Fig. 1). The MR damper is filled with MR fluid which is a composition of ferromagnetic particles suspended in carrier fluid. Characteristic feature of the piston are gaps which enable MR fluid to flow through it during damper compression and extension. MR fluid flowing in piston gaps is subjected to magnetic field which is induced by supplied built-in coils. When exposed to magnetic field ferromagnetic particles of MR fluid are polarized and form chain-like structure parallel to the field lines and perpendicular to the direction of flowing fluid. MR damper subjected to stress force axially causes flow of MR fluid in gaps which is counteracted by MR chains. It results in the increase of resultant damping force generated by MR damper. A gas accumulator consisting of a diaphragm covering pressured gas protects the damper from damage in case of its critical compression.

The paper is organized as follows. Section 2 presents a concise overview of MR damper models. Section 3 addresses a novel approach to MR damper hysteresis modelling using delay filters. In Section 4 experiment set-up is presented, results of MR damper model identification are reported and model validation is discussed.

2 Modelling of MR Damper Behaviour

In general, modelling of MR damper behaviour is focused on constructing such mathematical description that gives the possibility to predict damper force generated under different piston relative velocities and control cur-rents treated as excitations. Models are quite complicated since they are required to capture damper bilinear behaviour, hysteresis and saturation of the damping force.

Phenomenological models try to reflect MR damper behaviour by describing physical phenomena occurring during its operation. That involves mainly analysis of MR fluid flow through the damper piston gaps and influence of fluctuating magnetic field on MR fluid behaviour. In order to describe these phenomena Navier-Stokes equations accompanied with Maxwell's equations are utilized. Unfortunately, significant mathematical complexity of partial differential equations of these types makes this model impossible to be solved analytically. In addition, huge computational complexity of methods for numerical solution of partial differential equations makes such models difficult to incorporate into real-time applications. [4, 8]

In input-output models, an opposite approach is used, which does not concentrate on deeper understanding and describing physical phenomena appear-ing during MR damper operation but rather tries to reflect as accurately as possible the input-output interdependencies leaving behind physical meaning of parameters. In this group of models the following models were inves-tigated in the literature: well suited functions, polynomial, involution, fuzzy and neural models. [4, 9, 10] There also exists a wide range of heuristic models in which hyperbolic or cyclometric functions are used to describe input-output behaviour of MR dampers. [9, 11, 12]

Behavioural models try to take advantage of benefits related to both phenomenological and input-output models. They do not describe in a straightforward form physical phenomena occurring in the operation of MR damper but they are constructed taking into account the analogy in the behaviour of damper and a set of mechanical elements reflecting basic rheological properties such as elasticity, viscosity and plasticity. To obtain behavioural model these elements are interconnected in appropriate way taking into account their input-output behaviour as well as best fitting. In this group of models Bingham model, Gamoda-Filisko model, Visco-plastic Li model are considered. [5] Another group of models are based on Bouc-Wen model in which specially constructed nonlinear dynamic term modelling hysteretic behaviour of MR damper is introduced. [13] The most frequently used modifications of Bouc-Wen model are Spencer model and Yang model. [7, 14]

3 Modelling of MR Damper Force-Velocity Hysteretic Behaviour

Novel approach to the modelling of velocity-force hysteretic behaviour was presented in [15]. It involves application of a first order linear filter included into the model's velocity-force signal path. The other authors have located the first order linear filter at the output of the backbone shaping function block. In case of such model structure, the filter is used to process force signal which is strongly nonlinear due to backbone function and includes numerous harmonics, also for sinusoidal excitations. Such model structure [15] makes it difficult to find

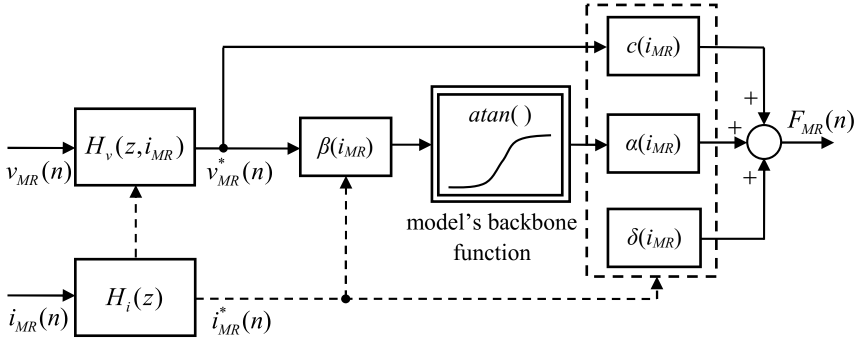


Fig. 2. Modelling of kinematic excitation and control dedicated signal paths of MR damper dynamics

appropriate phase shift parameters of the filter, which seems to be much simpler in case of the presented model.

In the current paper generalization of MR damper model, which is com-posed of dynamic and static components, is proposed (Fig. 2). Two kinematic excitation and control dedicated signal paths represent the dynamic part of the model. A backbone shaping function of the force-velocity characteristics is a static input-output component and may be represented by cyclometric (atan) or hyperbolic (tanh) functions.

Main component of presented model is backbone shaping function which is formed based on atan function presented in [12] as follows:

$$\begin{aligned} F_{MR}(n) = \alpha(i^*_{MR}) \cdot \text{atan}\{\beta(i^*_{MR}) \cdot v^*_{MR}(n) - \gamma \text{sign}[v^*_{MR}(n)]\}+ \\ + \delta(i^*_{MR}) + c(i^*_{MR}) \cdot v^*_{MR}(n), \end{aligned} \tag{1}$$

where v^*_{MR} is damper piston relative velocity obtained after filtering piston velocity v_{MR} using delay filters $H_v(z, i_{MR})$ included in velocity-force signal path. Parameters α, β, δ and c are to be estimated for different control current levels and excitation frequencies. Parameter γ describes hysteretic behaviour and in case of the presented model it equals zero.

Second part of the model consists of dynamic filters included in the velocity and control dedicated paths. The analysis is focused on dynamics exhibited by the velocity dedicated signal path which is modelled using all-pass delay filter $H_v(z, i_{MR})$ with flattened amplitude-frequency characteristics and v^*_{MR} is obtained as follows:

$$v^*_{MR}(n) = h_v(n, i_{MR}) * v_{MR}(n), \tag{2}$$

where $h_v(n, i_{MR})$ denotes the impulse reponse of the filter $H_v(z, i_{MR})$.

4 Results

Experiments results were obtained using Material Testing System (MTS) by courtesy of Department of Theoretical and Applied Mechanics, Silesian University of Technology. Peak-to-peak amplitude of damper piston rela-tive velocity excitation was assumed to be equal to 0.08 m/s. Experiments were performed for different kinematic excitation frequencies in range from 0.5 Hz to 2.5 Hz with resolution from 0.2 Hz to 0.5 Hz. Measurements were taken with sample rate of 1 kHz or 500 Hz in case of MTS or dedicated controller, respectively. Stabilization of current flowing through the damper coil was applied via PWM modulated voltage signal using PID algorithm with gain scheduling implemented in the controller. Additionally, temperature sensor was attached to the damper to measure temperature of damper housing. Force and displacement measurements performed by MTS were synchronized with control current and temperature measurements on the basis of acceleration measurements obtained from accelerometer attached to moving part of examined MR damper.

Specially generated current signal consisted of seven current levels cho-sen as 0.05, 0.1, 0.15, 0.25, 0.5, 0.75, 1.0 amperes to cover uniformly the range of forces generated by the MR damper. Current values were ordered nonli-nearly due to the nonlinear relationship between control current and dam-per force saturation. Time length dedicated to each experiment and each current level was assumed in such a way that for each current level at least 5 cycles of damper's sinusoidal excitation were performed.

4.1 Estimation of Time Delay Dedicated to Model Velocity Signal Path

Identification procedure was performed separately for each control cur-rent level and each frequency of velocity excitation. Cost function CF, which was minimized to determine model parameters, was defined as follows:

$$CF(\alpha, \beta, \gamma, c, N_d) = \frac{1}{N} \sum_{n} [F_{MR}(n) - \hat{F}_{MR}(n, \alpha, \beta, \gamma, \delta, c, N_d)]^2 \rightarrow \text{Max} \quad (3)$$

where:

$$\begin{aligned} \hat{F}_{MR}(n, \alpha, \beta, \delta, c, N_d) = {} & \alpha(i_{MR}) \cdot \text{atan}[\beta(i_{MR}) \cdot v_{MR}(n - N_d)] \\ & + \delta(i_{MR}) + c(i_{MR}) \cdot v_{MR}(n - N_d) \end{aligned} \quad (4)$$

and

$$v^*_{MR} v_{MR}(n - N_d). \quad (5)$$

Cost function CF was minimized by means of the *fmincon* function included in the Matlab Optimization Toolbox.

Estimated parameters were averaged with respect to the excitation frequency and are presented in Table 1 with respect to control current levels. It can be stated that parameters α and c are directly proportional to control current.

Table 1. Parameters of MR damper model related to current levels and fitting indices

Parameters of MR damper model using delay filters							
	i_{MR} [A]						
	0.05	0.10	0.15	0.25	0.50	0.75	1.0
α [N]	235.6	389.6	504.6	673.9	897.9	992.2	1056.9
β [ms^{-1}]	181.7	129.3	117.2	106.4	96.9	93.1	88.2
δ [N]	13.2	19.1	25.7	32.0	32.0	23.0	16.8
c [Nms^{-1}]	742.1	802.6	893.5	985.9	1279.0	1376.3	1367.0
Square root of Cost Function CF for:							
Delay filters based MR damper model							
	25.4	42.6	55.2	72.9	106.5	110.5	119.8
Acceleration based MR damper model [12]							
	25.1	42.4	55.0	73.1	107.5	112.6	122.8

Parameter β is inversely proportional to control current. Sensitivity of the model to offset parameter δ is the least in comparison to other parameters due to force zeroing procedure performed during each experiment. However, it is also included in the analysis.

Time delay values N_d estimated for each experiment condition (in Fig. 3 in the form of phase shift) can be used to synthesize bank of delay filters $H_v(z, i_{MR})$, each dedicated to a certain current level. It can be stated that the higher value of control current the greater time delay is exhibited by the velocity signal path. Moreover, phase shift characteristics are smooth and they decrease with respect to excitation frequency for all current levels.

High sampling rate dedicated to measurement data, analysis required to be performed for low frequencies and low values of phase shift make identification of delay filters parameters complex. Thus, synthesis of delay filters included in the presented model is not considered in the current paper.

4.2 Model Validation

Responses of the model were simulated for all experiment conditions. Some of them were graphically compared with measurement data. Results presented in Fig. 4 correspond to current levels equal to 0.05, 0.15, 0.5, 1.0 amperes and excitation frequency of 0.5, 0.9, 1.5 and 2.5 Hz, respectively. It can be noticed that level of saturated force generated by the damper depends directly on the current level. What is more, the higher excitation frequency, the wider is the hysteresis loop and such feature is also mapped by the presented model. More precise analysis of model fitting is performed using quality index defined as square root of the cost function (3)-(5) which is utilized in the identification procedure.

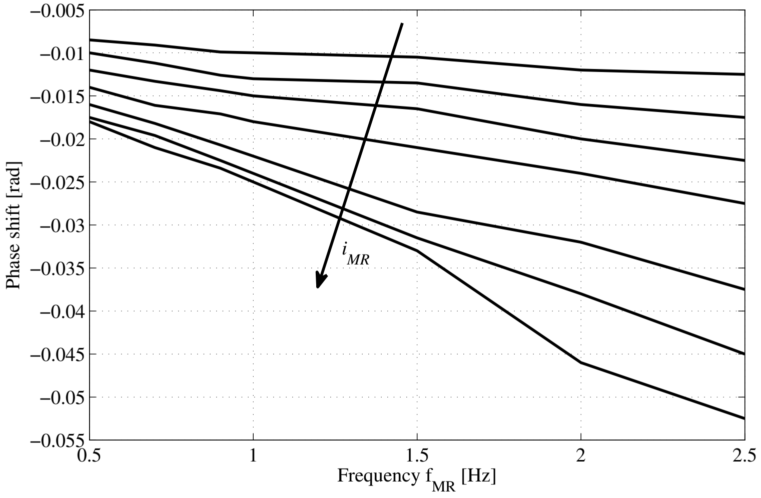


Fig. 3. Phase shift frequency characteristics of the desired velocity dedicated delay filters

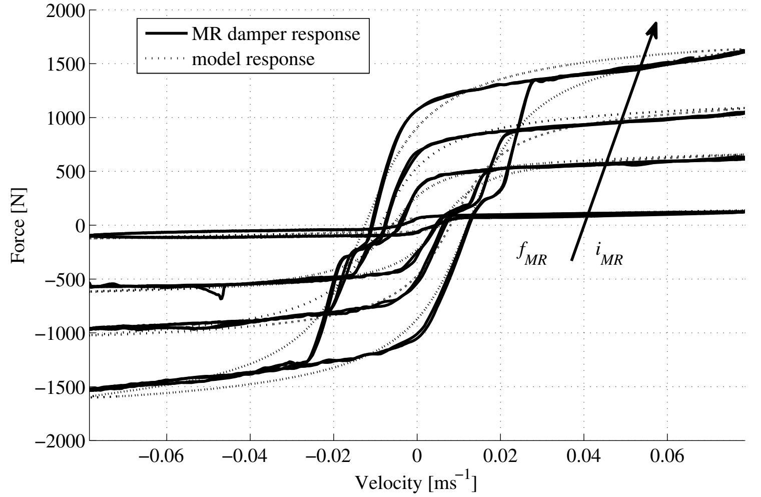


Fig. 4. Comparison of MR damper and model force response for different values of control current and frequency of piston velocity excitation

Apart from delay filter based model, also the acceleration based model [12] was identified and validated for all experiment conditions. For the purpose of quality examination, it should be also taken into account that in case of phenomenological hysteresis model the existence of phenomenological description is favoured over slightly better measurement fitness (acceleration based model [12]).

However, values of quality index which are listed and compared for both models in Table 1 justify at least comparable quality of fitting of model, which is based on delay filters, to measurement data.

5 Conclusions

Most models of MR damper which are well-known in literature possess quasi-static structure and do not take into account actual phenomena of damper behaviour. Hysteretic behaviour visible in MR damper's force-velocity characteristics is one of such features. Authors of the current paper presented a classification of MR damper's hysteresis models and those mostly known in literature were referred to as the input-output models. Authors claim that hysteretic behaviour of MR damper is caused by dynamics of velocity to force signal path which constitute a phenomenological model using delay filters. On the basis of experimental data, parameters of the presented model and reference quasi-static model were estimated and compared. Both models demonstrate comparable modelling accuracy. Moreover, presented frequency characteristics of desired delay indicate significant dependence on control current levels which needs to be taken into account in real-time applications.

Presented model will be utilized for the purpose of simulation and validation of semiactive vibration control algorithm. Moreover, future research will be also focused on complete synthesis of delay filters and their extension to other experimental conditions mainly corresponding to various amplitudes of excitation velocities.

Acknowledgements. The work reported in this paper has been partially financed by the National Science Centre, decision no. DEC-2011/01/B/ST7/06027.

References

1. Chen, Z.Q., Wang, X.Y., Ko, J., Spence Jr., B.F., Yang, G.: MR damping system on Dongt-ing Lake cable-stayed bridge. In: Proc. SPIE, vol. 5057, pp. 229–235 (2003)
2. Symans, M.D., Constantinou, M.C.: Semi-active control systems for seismic protection of structures: a state-of-the art review. J. Engineering Structures 21, 469–487 (1999)
3. Dong, X.M., Yu, M., Li, Z., Liao, C., Chen, W.: A comparison of suitable control methods for full vehicle with four MR dampers, part I: formulation of control schemes and numerical simulation. J. Intelligent Material Systems and Structures 20, 771–786 (2009)
4. Sapiński, B.: Magnetorheological dampers in vibration control. AGH University of Science and Technology Press, Cracow (2006)
5. Li, W.H., Yao, G.Z., Chen, G., Yeo, S.H., Yap, F.F.: Testing and steady state modeling of a linear MR damper under sinusoidal loading. J. Smart Materials and Structures 9, 95–102 (2000)

6. Choi, S.-B., Lee, S.-K., Park, Y.-P.: A hysteresis model for the field-dependent damping force of a magnetorheological damper. J. Sound and Vibration 245(2), 375–383 (2001)
7. Yang, S., Li, S., Wang, X., Gordaninejad, F., Hitchcock, G.: A hysteresis model for Magneto-rheological damper. Intern. J. Nonlinear Sciences and Numerical Simulation 6(2), 139–144 (2005)
8. Wereley, N.M., Pang, L.: Nondimentional analysis of semi-active electrorheological and magnetorheological dampers using approximate paralled plate models. J. Smart Materials and Structures 7, 732–743 (1998)
9. Plaza, K.: An empirical inverse magnetorheological damper model. In: Proc. International Conference on Methods and Models in Automation and Robotics, Międzyzdroje, Poland (2006)
10. Plaza, K.: True RMS-Based Inverse MR Damper Model for a Semi-Active System. In: Proc. International Conference on Methods and Models in Automation and Robotics, Międzyzdroje, Poland (in press, 2013)
11. Dalei, G., Haiyan, H.: Nonlinear stiffness of Magneto-Rheological Damper. J. Nonlinear Dynamics 40, 241–249 (2005)
12. Kasprzyk, J., Plaza, K., Wyrwal, J.: Identification of a magnetorheological damper for semi-active vibration control. In: Proc. International Congress on Sound and Vibration, Vilnius, Lithuania (2012)
13. Spencer Jr., B.F., Yang, G., Carlson, J.D., Sain, M.K.: Smart dampers for seismic protec-tion of structures: a full-scale study. In: Proc. 2nd World Conference on Structural Control (1998)
14. Spencer, B.F., Dyke, S.J., Sain, M.K., Carlson, J.D.: Phenomenological model of magnetor-hoelogical damper. J. of Eng. Mechanics, American Society of Civil Engineers 123(3), 230–238 (1997)
15. Song, X., Ahmadian, M., Southward, S.C.: Modeling Magnetorheological Dampers with Application of Nonparametric Approach. J. of Intelligent Material Systems and Structures 16, 421–432 (2005)

On Stability Criteria of Third-Order Autonomous Nonlinear Differential Equations with Quasi-Derivatives

Martin Neštický and Oleg Palumbíny

Institute of Applied Informatics and Mathematics,
Faculty of Materials Science and Technology,
Slovak University of Technology in Bratislava,
Hajdoczyho 1, 917 24 Trnava, Slovakia
{martin.nesticky,oleg.palumbiny}@stuba.sk

Abstract. The paper deals with ordinary 3-order nonlinear differential equations $L_3y = f(L_0y, L_1y, L_2y)$ with quasi-derivatives. There is established a criterion of asymptotic stability in Liapunov sense as well as a criterion of instability in Liapunov sense. The results are illustrated by proper examples.

Keywords: Nonlinear differential equation, 3-rd order, quasi-derivative, stability in Liapunov sense, asymptotic stability in Liapunov sense, instability in Liapunov sense.

1 Introduction

There are many articles from Control Theory concerning the stability criteria of nonlinear differential equations with classical derivatives. In the paper [5] Palumbíny derived stability criteria of certain class of third-order nonlinear differential equations with so called quasi-derivatives (see the text below). The main aim of the presented article is to establish similar criteria for another class of third-order nonlinear differential equations with quasi-derivatives, so called *autonomous equations*. It means that the function f does not explicitely depend on the independent variable t.

Our paper deals with a criterion of asymptotic stability as well as instability, both in Liapunov sense, of a null solution 0 of the autonomous nonlinear third-order differential equations with the quasi-derivatives

$$\text{(L)} \qquad L_3y = f(L_0y, L_1y, L_2y)$$

where (the prime means a derivative owing to the variable t)

$$\begin{aligned} L_0y(t) &= y(t), \\ L_1y(t) &= p_1(t)\,(L_0y(t))', \\ L_2y(t) &= p_2(t)\,(L_1y(t))', \\ L_3y(t) &= (L_2y(t))', \end{aligned}$$

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_42, © Springer International Publishing Switzerland 2014

$p_i(t)$, $i = 1, 2$ are real-valued continuous functions defined on an opened real interval (b, ∞), $f(y_1, y_2, y_3)$ is real-valued and continuous up to all 2-nd order partial derivatives of the function f on some area $H \subset E^3$, $\mathbf{o} \in H$ where the symbol E denotes the set of all real numbers. The symbol $\mathbf{o}$ means the vector $(0, 0, 0)$. *The null solution* 0 is a function equals to zero for all t such this null function is a solution (L) on (b, ∞). The terms $L_k y(t)$, $k = 0, 1, 2, 3$ are k-th *quasi-derivatives* of a function $y(t)$.

We recall that function f does not explicitly depend on the independent variable t. We note that we shall use a matrix norm of the form $\|\{a_{ij}\}_{i,j}\| = \sum_{i,j} |a_{ij}|$. A set [–1, 1] is a closed interval with bounds –1, 1.

An important contribution of the article consist in the fact that there are more control parametres in (L). These parametres are functions $p_i(t)$, $i = 1, 2$ which enable an user a better control of considered processes described by the differential equation (L).

Remark 1. The differential equation (L) can be expressed more detailed as

$$\text{(M)} \qquad \left(p_2(t)\left(p_1(t)y'\right)'\right)' = f\left(y, p_1(t)y', p_2(t)\left(p_1(t)y'\right)'\right).$$

Let us consider a differential system of the first order

$$\text{(S)} \qquad \begin{aligned} y_1' &= f_1(t, y_1, y_2, y_3) \\ y_2' &= f_2(t, y_1, y_2, y_3) \\ y_3' &= f_3(t, y_1, y_2, y_3). \end{aligned}$$

Assumption. Let the system (S) be expressed in a matrix form $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$. Through out the paper we shall assume an existence of a number b (real or $-\infty$) and an area $H \subset E_1^3$, $\mathbf{o} \in H$ such that the function $\mathbf{f}$ is continuous on $G = (b, \infty) \times H$ and for every point $(\tau, \mathbf{k}) \in G$ the following Cauchy problem

$$\text{(1)} \qquad \mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(\tau) = \mathbf{k},$$

admits the only solution. We also assume $\mathbf{f}(t, \mathbf{o}) = \mathbf{o}$ for all $t > b$, i.e. the Cauchy problem (1) admits for $\mathbf{k} = \mathbf{o}$ the trivial solution defined by a formula $\mathbf{o}(t) = \mathbf{o}$ for all $t > b$.

Definition 1. We say that *the trivial solution* $\mathbf{o}$ of the system (S) *is asymptotically stable in Liapunov sense*, if for every $\tau > b$ and every $\epsilon > 0$ there exists $\delta = \delta(\tau, \epsilon) > 0$ such that for every initial values $\mathbf{k} \in H$, $\|\mathbf{k}\| < \delta$ and for all $t \geq \tau$ it holds that the solution $\mathbf{u}(t, \tau, \mathbf{k})$ of the Cauchy problem (1) fulfils the following inequality

$$\text{(2)} \qquad \|\mathbf{u}(t, \tau, \mathbf{k})\| < \epsilon.$$

Otherwise, *the trivial solution* $\mathbf{o}$ *is instable in Liapunov sense.*

Definition 2. We say that *the trivial solution* $\mathbf{o}$ of the system (S) *is asymptotically stable in Liapunov sense*, if the trivial solution $\mathbf{o}$ is stable in Liapunov sense and there exists a real number $\varDelta > 0$ such that for all $\mathbf{k} \in H$, $\|\mathbf{k}\| < \varDelta$ and for every $\tau > b$ it holds that

$$\lim_{t\to\infty} \|\mathbf{u}(t,\tau,\mathbf{k})\| = 0.$$

Definition 3. A special type (T) of the system (S) of the form

$$\text{(T)} \qquad \begin{aligned} y_1' &= y_2/p_1(t) \\ y_2' &= y_3/p_2(t) \\ y_3' &= f(y_1,y_2,y_3) \end{aligned}$$

is called *a competent system to the equation* (L).

Remark 2. We recall an important property of the system (T) which consist in a fact that a function $u(t)$ is a solution of (L) if and only if a vector $(u(t), L_1u(t), L_2u(t))$ is a solution of (T).

Definition 4. Let (L) be such an equation that the function 0 is a solution of (L) on (b,∞). Then, according to Remark 2, the vector (0,0,0) is a solution of (T). We say 0 is *a stable solution of* (L) *in Liapunov sense,* if (0,0,0) is a stable solution of (T) in Liapunov sense. Otherwise, 0 is *an instable solution of* (L) *in Liapunov sense.*

Definition 5. Let (L) be such an equation that 0 is a solution of (L) on (b,∞). Then, according to Remark 2, the vector (0,0,0) is a solution of (T). We say 0 is *an asymptotically stable solution of* (L) *in Liapunov sense*, if (0,0,0) is an asymptotically stable solution of (T) in Liapunov sense.

The main aim of the paper is to establish the criteria, which assure the asymptotic stability as well as instability of the null solution 0 of the equation (L). If we put $p_k(t) = 1$ on (b,∞), $k = 1,2$ in (L), we obtain a differential equation with classic derivatives. We note that the functions $p_k(t)$, $k = 1,2$ are not, in general, assumed to be differentiable. From this it follows that we cannot use on (L) stability criteria derived for nonlinear differential equations with classic derivatives.

2 Auxiliary Assertions

Now we introduce some auxiliary assertions, which are significant according our considerations. The first of them is the special case of the wellknown Hurwitz criterion when $n = 3$ (see[2], Chapter 9):

Theorem 1. Let us consider a polynomial

$$\text{(3)} \qquad b_3s^3 + b_2s^2 + b_1s + b_0,$$

where b_i, $i = 0, 1, 2, 3$ are real numbers such that $b_0 > 0$, $b_3 \neq 0$. Then all zeros of the polynomial (3) admit negative real parts if and only if it holds that

$$b_1 > 0, \tag{4}$$

$$b_1 b_2 - b_0 b_3 > 0, \tag{5}$$

$$b_1 b_2 b_3 - b_0 b_3^2 > 0. \tag{6}$$

The second assertion deals with asymptotic criteria of stability in Liapunov sense for systems of differential equations of the first order (see [1], Chapter 13 or [4]):

Theorem 2. Let us consider a system of differential equations of the first order expressed in the following matrix form

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{B}(t)\mathbf{x} + \mathbf{g}(t, \mathbf{x}), \qquad \mathbf{g}(t, \mathbf{o}) = \mathbf{o}, \tag{7}$$

where $\mathbf{A}$ is a real constant square matrix, $\mathbf{B}(t)$ is a real square matrix depending on t only, such that

$$\lim_{t\to\infty} \mathbf{B}(t) = \mathbf{0}, \tag{8}$$

where $\mathbf{0}$ is the null matrix and $\mathbf{g}$ a real vector function continuous on an area $(b, \infty) \times H$, where $b \in E$, satisfying a condition

$$\lim_{\|\mathbf{x}\|\to 0} \frac{\|\mathbf{g}(t, \mathbf{x})\|}{\|\mathbf{x}\|} = 0 \text{ uniformly for all } t \geq b. \tag{9}$$

Then:

(i) If all eigenvalues of $\mathbf{A}$ have negative real parts, then the trivial solution of (7) is asymptotically stable in Liapunov sense.

(ii) If at least one of eigenvalues of $\mathbf{A}$ has a positive real part, then the trivial solution of (7) is instable in Liapunov sense.

3 Results

Now we shall prove the first main result of the paper – the criterion of asymptotic stability of the null solution 0 in Liapunov sense of the differential equation (L):

Theorem 3. Let us consider the differential equation (L) such that

(a) $\lim_{t\to\infty} p_i(t) = a_i > 0$, $i = 1, 2$,

If it holds that

$$\text{(b)} \qquad \frac{\partial f}{\partial w_1}(\mathbf{o}) < 0, \quad \frac{\partial f}{\partial w_3}(\mathbf{o}) < 0, \quad a_1 \frac{\partial f}{\partial w_2}(\mathbf{o}) \frac{\partial f}{\partial w_3}(\mathbf{o}) + \frac{\partial f}{\partial w_1}(\mathbf{o}) > 0,$$

then the null solution 0 of (L) on (b, ∞) is asymptotically stable in Liapunov sense.

Proof. The null solution 0 is, according to Definition 5, asymptotically stable in Liapunov sense, if the solution $\mathbf{o}$ of the system (T) is asymptotically stable in Liapunov sense, where (T) is expressed in the form (U), where

$$\text{(U)} \qquad \mathbf{w}' = \mathbf{A}\mathbf{w} + \mathbf{B}(t)\mathbf{w} + \mathbf{g}(t, \mathbf{w}), \qquad \mathbf{g}(t, \mathbf{o}) = \mathbf{o},$$

and

$$\mathbf{w}' = \begin{bmatrix} w_1' \\ w_2' \\ w_3' \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}, \quad \mathbf{o} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0, & \dfrac{1}{a_1}, & 0, \\ 0, & 0, & \dfrac{1}{a_2}, \\ \dfrac{\partial f}{\partial w_1}(\mathbf{o}), & \dfrac{\partial f}{\partial w_2}(\mathbf{o}), & \dfrac{\partial f}{\partial w_3}(\mathbf{o}), \end{bmatrix},$$

$$\mathbf{B}(t) = \begin{bmatrix} 0, & \dfrac{1}{p_1(t)} - \dfrac{1}{a_1}, & 0, \\ 0, & 0, & \dfrac{1}{p_2(t)} - \dfrac{1}{a_2}, \\ 0, & 0, & 0, \end{bmatrix}, \qquad \mathbf{g}(t, \mathbf{w}) = \begin{bmatrix} 0 \\ 0 \\ \dfrac{1}{2}\mathrm{d}^2 f(\theta \mathbf{w}, \mathbf{o}) \end{bmatrix},$$

$$\begin{aligned} \frac{1}{2}\mathrm{d}^2 f(\theta \mathbf{w}, \mathbf{o}) = {} & \frac{1}{2}\frac{\partial^2 f}{\partial w_1^2}(\theta \mathbf{w}) w_1^2 + \frac{1}{2}\frac{\partial^2 f}{\partial w_2^2}(\theta \mathbf{w}) w_2^2 + \frac{1}{2}\frac{\partial^2 f}{\partial w_3^2}(\theta \mathbf{w}) w_3^2 + \\ & + \frac{\partial^2 f}{\partial w_1 w_2}(\theta \mathbf{w}) w_1 w_2 + \frac{\partial^2 f}{\partial w_1 w_3}(\theta \mathbf{w}) w_1 w_3 + \frac{\partial^2 f}{\partial w_2 w_3}(\theta \mathbf{w}) w_2 w_3 \end{aligned}$$

where we used the Taylor's theorem with the remainder in the Lagrange's form as $k = 2, \ 0 < \theta < 1$. Then

$$\begin{aligned} 0 \le \frac{\|\mathbf{g}(t, \mathbf{w})\|}{\|\mathbf{w}\|} &= \frac{|0| + |0| + |\frac{1}{2}\mathrm{d}^2 f(\theta \mathbf{w}, \mathbf{o})|}{|w_1| + |w_2| + |w_3|} \le \frac{1}{|w_1| + |w_2| + |w_3|} \times \\ & \times \left(\left| \frac{1}{2}\frac{\partial^2 f}{\partial w_1^2}(\theta \mathbf{w}) w_1^2 \right| + \left| \frac{1}{2}\frac{\partial^2 f}{\partial w_2^2}(\theta | \mathbf{w}) w_2^2 \right| + \left| \frac{1}{2}\frac{\partial^2 f}{\partial w_3^2}(\theta \mathbf{w}) w_3^2 \right| + \right. \\ & \left. + \left| \frac{\partial^2 f}{\partial w_1 w_2}(\theta \mathbf{w}) w_1 w_2 \right| + \left| \frac{\partial^2 f}{\partial w_1 w_3}(\theta \mathbf{w}) w_1 w_3 \right| + \left| \frac{\partial^2 f}{\partial w_2 w_3}(\theta \mathbf{w}) w_2 w_3 \right| \right) \end{aligned}$$

Without lost of generality, we assume that $(w_1, w_2, w_3) \in [-1,1]^3$. From this and from the continuity of all 2-nd order partial derivatives of the function f it follows that exists a positive real constant K such that

$$\left|\frac{1}{2}\frac{\partial^2 f}{\partial w_k^2}(\theta\mathbf{w})\right| \leq K,\ k = 1,2,3,$$

$$\left|\frac{\partial^2 f}{\partial w_i w_j}(\theta\mathbf{w})\right| \leq K,\ (i,j) = (1,2),(1,3),(2,3) \tag{11}$$

for all $(w_1, w_2, w_3) \in [-1,1]^3$. If we use the estimations (11) in the formula (10), we obtain

$$0 \leq \frac{\|\mathbf{g}(t,\mathbf{w})\|}{\|\mathbf{w}\|} = \frac{|0| + |0| + |\frac{1}{2}\mathrm{d}^2 f(\theta\mathbf{w},\mathbf{o})|}{|w_1| + |w_2| + |w_3|} \leq \frac{K}{|w_1| + |w_2| + |w_3|} \times$$
$$\times \left(|w_1^2| + |w_2^2| + |w_3^2| + |w_1 w_2| + |w_1 w_3| + |w_2 w_3|\right) =$$
$$= K\left(\frac{|w_1| + |w_2|}{|w_1| + |w_2| + |w_3|}|w_1| + \frac{|w_2| + |w_3|}{|w_1| + |w_2| + |w_3|}|w_2| + \right.$$
$$\left. + \frac{|w_1| + |w_3|}{|w_1| + |w_2| + |w_3|}|w_3|\right) \leq K\left(|w_1| + |w_2| + |w_3|\right) = K\|\mathbf{w}\|.$$

The squeeze theorem from Limit Theory yields that $\dfrac{\|\mathbf{g}(t,\mathbf{w})\|}{\|\mathbf{w}\|}$ converges to zero as $\|\mathbf{w}\| \to 0$. This convergence is uniform because the term $K\left(|w_1| + |w_2| + |w_3|\right)$ does not explicitly depend on the variable t. From this it implies that (9) hold. We can easily observe a validity of the conditions (7), (8) in Theorem 2. The validity of condition (a), (b) assure that the conditions (4),(5),(6) hold in Theorem 1, where the characteristic polynomial of the matrix $\mathbf{A}$ is

$$s^3 - \frac{\partial f}{\partial w_3}(\mathbf{o})s^2 - \frac{1}{a_2}\frac{\partial f}{\partial w_2}(\mathbf{o})s - \frac{1}{a_1 a_2}\frac{\partial f}{\partial w_1}(\mathbf{o}).$$

Then Theorem 1 yields that all the eigenvalues of $\mathbf{A}$ have the negative real parts. Consequently, Theorem 2, the part (i) as well as Definition 5 yield the required stability of null solution 0 of (L). □

Example 1. Let us consider the nonlinear differential equation (L), where $p_1(t) = 2 + \dfrac{1}{t}, p_2(t) = 3 - \dfrac{1}{t}$ and

$$L_3 y = \frac{y^4}{1 + (L_2 y)^2} - y - 2L_1 y - 3L_2 y.$$

It is obvious that Assumption, mentioned after Remark 1, hold for $b = 0$. An easy computing yield that (L) admits the null solution 0 as well as $a_1 = 2$,

$a_2 = 3, \dfrac{\partial f}{\partial w_1}(\mathbf{o}) = -1, \dfrac{\partial f}{\partial w_2}(\mathbf{o}) = -2, \dfrac{\partial f}{\partial w_3}(\mathbf{o}) = -3$. From this immediately follows the validity of (a) and (b) in Theorem 3. Then the last mentioned theorem yields required stability of the null solution 0 of the equation (L).

Now we shall prove the second main result of the paper – the criterion of instability of the null solution 0 in Liapunov sense of the differential equation (L):

Theorem 4. Let us consider the differential equation (L) such that

(a) $\lim\limits_{t\to\infty} p_i(t) = a_i > 0$, $i = 1, 2$.

If it holds that

(b') at least one real part of zeros of

$$s^3 - \frac{\partial f}{\partial w_3}(\mathbf{o})s^2 - \frac{1}{a_2}\frac{\partial f}{\partial w_2}(\mathbf{o})s - \frac{1}{a_1 a_2}\frac{\partial f}{\partial w_1}(\mathbf{o}) \text{ is positive,}$$

then the null solution 0 of the equation (L) is instable in Liapunov sense.

Proof. The null solution 0 of (L) is, according to Definition 4, instable in Liapunov sense, if the solution (0,0,0) of the system (U) is instable in Liapunov sense. By the same way as in the proof of Theorem 3, it can by proved the validity of (9). We can easily observe the validity of the condition (7), (8) in Theorem 2. Then the last mentioned Theorem, the part (ii) yields the required instability of the null solution 0 of (L) □

Example 2. Let us consider the equation (L) where $p_1(t) = 2 + \dfrac{1}{t}, p_2(t) = 3 - \dfrac{1}{t}$ and

$$L_3 y = \frac{y^6}{4 + (L_1 y)^2} + y - 2L_1 y - 3L_2 y.$$

It is obvious that Assumption, mentioned after Remark 1, hold for $b = 0$. An easy computing yields that the function 0 is a solution of (L) as well as $a_1 = 2$, $a_2 = 3, \dfrac{\partial f}{\partial w_1}(\mathbf{o}) = 1, \dfrac{\partial f}{\partial w_2}(\mathbf{o}) = -2, \dfrac{\partial f}{\partial w_3}(\mathbf{o}) = -3$. From this immediately follows the validity of (a). By the same way as in Example 1 it can be proved (b'). The characteristic polynomial of $\mathbf{A}$ is

$$h(s) = s^3 + 3s^2 + \frac{2}{3}s - \frac{1}{6}.$$

There are two possibilities only: 1) $h(s)$ admits three real zeros. Then their product is equal to 1/6. It means, that all this zeros differ null. If all these zeros were negative, then their product would be negative, which is a contradiction. 2) $h(s)$ admits one real zero a and two complex zeros $b \pm ci$. Then their product $a(b^2 + c^2)$ equals to 1/6 again. If $a \leq 0$, than this product would be nonpositive. Thus $a > 0$. Then, owing to Theorem 4, the function 0 is an instable solution of (L) in Liapunov sense.

4 Conclusions

The foregoing results can be used for ordinary nonlinear differential equations with quasi-derivatives. The differential equations in applications where the quasi-derivatives have been occured are, for example, the differential equations describing a stationary distribution of temperature in a wall of a circle tube as well as the differential equations of an equilibrium state of a straight mass bar. For more details see [6].

References

1. King, A.C.: Differential Equations. Cambridge University Press, Birmingham (2003)
2. Teschl, G.: Ordinary Differential Equations and Dynamical Systems. AMS Publ., Rhode Island (2012)
3. Zabczyk, J.: Mathematical Control Theory. Birkhauser, Boston (2008)
4. Nagy, J.: Stabilita rieseni obycajnych diferencialnych rovnic. In: MVST XVI, SNTL Praha (1980)
5. Palumbíny, O.: On Stability Criteria of Third Order Nonlinear Differential Equations. In: Second International Conferences on Electronic, Communication and Control (2012)
6. Kneschke, A.: Differentialgleichungen und Randwertprobleme, vol. 3. B. G. Teubner, Leipzig (1968)

An M/G/1 Retrial Queue with Working Vacation

Amar Aissani[1], Samira Taleb[1], Tewfik Kernane[1],
Ghania Saidi[2], and Djamel Hamadouche[3]

[1] University of Science and Technology Houari Boumediene (USTHB),
BP 32 El Alia, Bab-Ez-Zouar. 16 111, Algiers, Algeria
aaissani@usthb.dz, talebsamira04@yahoo.fr, tkernane@gmail.com
[2] High School of Statistics and Applied Economics ESSEA,
11 chemin Doudou Mokhtar, Ben Aknoun. Algiers, Algeria
ghsaidi@yahoo.fr
[3] Mouloud Mammeri University of Tizi-Ouzou (UMMTO),
Site Bastos,Tizi-Ouzou, Algeria
djhamad@ummto.dz

Abstract. In this paper, we show through the example of the M/G/1 queue with working vacations, how queueing theory can help to the performance evaluation of some modern systems. We obtain the joint probability distribution of the server state and the number of orbiting customers in the system.This distribution is obtained in terms of Laplace and z- transforms. We show how mean performance measures can be obtained.

Keywords: Mass Service, Retrial queues, Working vacations, Piece-Wise Markov Process, Laplace and z transforms.

1 Introduction

A Queueing (or a mass service) model of a complex system is a formal model in which the server represents access of customers to resources and queue capacity models, resource restrictions and storage before service, with some queueing discipline (FIFO,LIFO, RANDOM, Processor Sharing,). This theory have been developed from different point of views in Operation Research, Applied Probability and Computer Science motivated by the progress of computer technologies and networks in many areas such as telecommunication, flexible manufacturing, supply-chain, e-commerce.

In practical situations, the server can be subject to some interruptions (priorities or breakdowns:

- (i) random interruptions (which are qualified as breakdowns). These interruptions of service can be due to random events: physical (mechanical or electronic one's) or software nature (bug in the program or attack).

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_43, © Springer International Publishing Switzerland 2014

- (ii) programmed interruptions (called "vacations"), which allows to exploit the idle time for secondary tasks (preventive maintenance, priority jobs, security actions..). Such a policy allows more flexibility (for example, digital nomadism or connected mobility) in the the optimal design and control of a queueing system or network from the point of view of the QoS (Quality of Service).

Now, in order to manage these (random) situations, we can consider different "control" options [4]

- (i) the server can be turned off and takes a vacation of random length whenever the system is empty.
- (ii) the server is turned on when the accumulation of units in the system is greater than a fixed threshold N or after a fixed period of time T.
- (iii) the control policy may allow a single vacation (and then waiting for a new requests) or multiple vacations (so, the server takes vacations until he finds at least one request in the file).

Some recent contributions gave orientations to a new type of vacation, namely working vacations. By working vacation, Servi & Finn [7] means that a single server works at a different rate rather than completely stopping during the vacation period. This work try to model (using the $M/M/1$ version) wavelength division multiplexing optical access network (WDN) using multiple wavelengths which can be reconfigured. Another application concerns the fact that the trade-off between benefit of working on other jobs and cost of increasing waiting time of the queue can be achieved by designing the appropriate vacation policy. The system with Markov Arrivals Process $MAP/G/1$ is considered in [11] and the model with balk arrivals in [9]. The interested reader can found more references in the survey by Ke & al [4].

In classical queueing models, an arrival finding the server blocked (busy or out of order) joins a queue with some service discipline (FIFO,LIFO,RANDOM,), or it is considered to be lost unit (Erlang model). If the server is unavailable, the arrival can join a retrial group(or an "orbit" , which is a sort of queue for secondary sources) and repeat successively an attempt until the server is able to provide service. Otherwise, if the server is available, the arriving request begins service immediately. Such models are called systems with repeated calls or retrial queues, see for example the accessible bibliography of Artalejo [1].The retrial $M/M/1$ version with working vacations has been studied by Do [2] where during a vacation the customer is served with a constant rate (so, an exponential service distribution) smaller than the normal service rate.The $M/G/1$ version with exponential vacations has been investigated for example by Li & al [5].

In this work, we consider a model which combines both vacation and retrial phenomena. In this work, we consider an extension of the M/G/1 retrial queue. During the vacation period, the service time is smaller (or greater) stochastically (or in distribution) than the service time in normal period.In the following section we describe the mathematical model. The section 3 gives the ergodicity condition. In section 4 we derive a system of differential equations for the steady-state joint probability distribution of the server state and the number of customers in orbit. Section 5 shows how main performance metrics can be obtained.

2 The Mathematical Model

We consider an M/G/1 retrial queue with working vacations. The inter-arrival times of primary requests are exponentially distributed with parameter $\lambda > 0$, so the number of primary requests is a Poisson process with the same parameter. There is no queue in the classical sense. If an arriving primary call finds the server available and free of service, it immediately occupies the server and leaves the system after completion of service. If an arriving primary call finds the server blocked (occupied by a service or in vacation), it becomes a source of secondary call and return later to try again until it finds the server free and available; the collection of all secondary calls is called "orbit" (a sort of queue). Request retrials from the orbit of infinite size follow a Poisson process with constant rate ν [1].

The server takes a working vacation when at a service completion epoch the server is idle and the orbit empty. Vacation durations are exponentially distributed with parameter θ.The service times of the customers in normal mode form a sequence $\{S_b^n, n \geq 1\}$ of independent and identically distributed (i.i.d.) random variables with common probability distribution function $H_b(x), H_b(0+) = 0$ and Laplace-Stieltjes transform $h_b(s), Re(s) \geq 0$; first order moments are denoted by h_{1b} and h_{2b}.The service times of the customers who finds the server in vacation are i.i.d. random variables $\{S_v^n, n \geq 1\}$ with common probability distribution function $H_v(x), H_v(0+) = 0$ and Laplace-Stieltjes transform $h_v(s), Re(s) \geq 0$; first order moments are denoted by h_{1v} and h_{2v}.Let S_b and S_v be the generic service times in normal mode and during vacation respectively. We assume that $S_v \leq_{st} S_b$ that is the service during vacation is stochastically smaller than the service in normal mode. At the end of a vacation, the server takes another vacation if the service is idle and the orbit empty.

At the end of each vacation, the server only takes another vacation if there is no any new request or repeated request from the orbit.

Consider the following random process $\zeta(t) = \{\alpha(t), \beta(t), R(t); \xi(t), t \geq 0\}$, where $\{R(t), t \geq 0\}$ is the number of customers in orbit at time t; $\alpha(t) = 0$, if the server is idle and $\alpha(t) = 1$ if it is busy by the service of some customer; $\beta(t) = 0$, if the server is not on vacation and $\beta(t) = 1$ if it is on vacation.

We introduce the real positive random variable $\xi(t)$ which represents the residual service time in normal or vacation service modes if $\alpha(t) = 1$.

It is not difficult to show that the stochastic process $\{\zeta(t), t \geq 0\}$ is a Markovian process with piecewise linear paths which describes the evolution of the server state and the number of orbiting customers. We establish first the ergodicity condition for such a process, then we obtain its stationary probability distribution.

3 Ergodicity Condition

The following theorem gives a condition for the existence of a stationary regime.

Theorem 1. *The stochastic process* $\{\zeta(t), t \geq 0\}$ *is ergodic if the following condition holds:*

$$\frac{\lambda+\nu}{\nu}\tau_b < 1 \; , \qquad (1)$$

$$\frac{(\lambda+\nu)h_v(\theta)}{\lambda+\nu+\theta} < 1 \; . \qquad (2)$$

Proof. An heuristic proof of the sufficiency is provided in appendix B.

4 Joint Distribution of the Server State and the Number of Customers in Orbit

In this section we derive the joint distribution of the server state and the number of customers in orbit in steady-state by it's transform. Under the assumptions 1 and 2, the stochastic process $\{\zeta(t), t \geq 0\}$ is ergodic. As a consequence, the ergodic stationary probabilities

$$P_{ij}(m) = lim_{t\to\infty}P\{\alpha(t)=i, \beta(t)=j, R(t)=m\}, m \geq 0, (i,j)=(0,0),(0,1) \; ,$$

$$P_{ij}(m,x) = lim_{t\to\infty}P\{\alpha(t)=i, \beta(t)=j, R(t)=m; \xi(t)<x\},$$
$$i,j=(1,0),(1,1), m \geq 0, x \geq 0 \; .$$

are solutions of the following system of differential equations

$$P_{00}(0) \equiv 0 \; ,$$

$$(\lambda+\nu)P_{00}(m) = \frac{dP_{10}(m,0)}{dx} + \theta P_{01}(m), m \geq 1 \; ,$$

$$(\lambda+\theta)P_{01}(0) = \frac{dP_{11}(0,0)}{dx} + \frac{dP_{10}(0,0)}{dx} \; ,$$

$$(\lambda+\theta+\nu)P_{01}(m) = \frac{dP_{11}(m,0)}{dx}, m \geq 1 \; ,$$

$$\lambda P_{10}(0,x) = \frac{dP_{10}(0,x)}{dx} - \frac{dP_{10}(0,0)}{dx} + \nu P_{00}(1)H_b(x) + \theta P_{11}(0,x) \; ,$$

$$\lambda P_{10}(m,x) = \frac{dP_{10}(m,x)}{dx} - \frac{dP_{10}(m,0)}{dx} + \lambda P_{00}(m)H_b(x) +$$
$$+\nu P_{00}(m+1)H_b(x) + \theta P_{11}(m,x) + \lambda P_{10}(m-1,x), m \geq 1 \; ,$$

$$(\lambda+\theta)P_{11}(0,x) = \frac{dP_{11}(0,x)}{dx} - \frac{dP_{11}(0,0)}{dx} +$$
$$+\nu P_{01}(1)H_v(x) + \lambda P_{01}(0)H_v(x) \; ,$$

$$(\lambda+\theta)P_{11}(m,x)=\frac{dP_{11}(m,x)}{dx}-\frac{dP_{11}(m,0)}{dx}+$$
$$+\lambda P_{11}(m-1,x)+\nu P_{01}(m+1)H_v(x)+\lambda P_{01}(m)H_v(x), m\geq 1\ .$$

We introduce the partial generating functions in z,

$$Q_{ij}(z)=\sum_{m=0}^{\infty}z^m P_{ij}(m),(i,j)=(0,0),(0,1)\ ,$$

$$Q_{ij}(z,x)=\sum_{m=0}^{\infty}z^m P_{ij}(m,x),(i,j)=(1,0),(1,1)\ .$$

Applying these transforms to the previous system, we obtain

$$(\lambda+\nu)Q_{00}(z)=\frac{\partial Q_{10}(z,0)}{\partial x}+\theta Q_{01}(z)-\theta P_{01}(0)\ , \tag{3}$$

$$(\lambda+\theta+\nu)Q_{01}(z)-\nu P_{01}(0)=\frac{\partial Q_{11}(z,0)}{\partial x}+\frac{dP_{10}(0,0)}{dx}\ , \tag{4}$$

$$(\lambda-\lambda z)Q_{10}(z,x)=\frac{\partial Q_{10}(z,x)}{\partial x}-\frac{\partial Q_{10}(z,0)}{\partial x}+\lambda Q_{00}(z)H_b(x)+$$
$$+\frac{\nu}{z}Q_{00}(z)H_b(x)+\theta Q_{11}(z,x)\ , \tag{5}$$

$$(\lambda-\lambda z+\theta)Q_{11}(z,x)=\frac{\partial Q_{11}(z,x)}{\partial x}-\frac{\partial Q_{11}(z,0)}{\partial x}+\lambda Q_{01}(z)H_v(x)+\frac{\nu}{z}Q_{00}(z)H_v(x)\ . \tag{6}$$

Consider now the Laplace transform of the partial generating functions $Q_{ij}(z,x)$:

$$f_{ij}(z,s)=\int_0^{\infty}e^{-sx}Q_{ij}(z,x)dx,(i,j)=(1,0),(1,1)\ ,$$

and the Laplace-Stieltjes transforms

$$h_b(s)=\int_0^{\infty}e^{-sx}dH_b(x)\ ,$$

$$h_v(s)=\int_0^{\infty}e^{-sx}dH_v(x)\ .$$

We apply now the Laplace transform to the second argument of the obtained partial generating functions 5 and 6

$$s(s-\lambda+\lambda z)f_{10}(z,s)=\frac{\partial Q_{10}(z,0)}{\partial x}-\Big(\lambda+\frac{\nu}{z}\Big)Q_{00}(z)h_b(s)-\theta sf_{11}(z,s)\ , \tag{7}$$

$$s(s-\lambda+\lambda z-\theta)f_{11}(z,s)=\frac{\partial Q_{11}(z,0)}{\partial x}-\Big(\lambda+\frac{\nu}{z}\Big)Q_{01}(z)h_v(s)\ . \tag{8}$$

There is now some uncertainty about the unknown functions $\frac{\partial Q_{ij}(z,0)}{\partial x}$ which can be determined as usual by using the fact that these functions are analytical functions in the domain $||z|| \leq 1$, $Re(s) \geq 0$. Consider for example the first equation of the previous system of equations. Since $f_{11}(z,s)$ is analytic in the domain $Re(s) \geq 0$, and since the left right hand side of 8 is equal to zero for $s = \lambda - \lambda z + \theta$, then the right hand side must also be zero at this point.So, we have the first condition that

$$\frac{\partial Q_{11}(z,0)}{\partial x} = \Big(\lambda + \frac{\nu}{z}\Big) Q_{01}(z) h_v(\lambda - \lambda z + \theta) \; . \tag{9}$$

Thus,

$$f_{11}(z,s) = \Big(\lambda + \frac{\nu}{z}\Big) Q_{01}(z) \frac{h_v(\lambda - \lambda z + \theta) - h_v(s)}{s(s - \lambda + \lambda z - \theta)} \; . \tag{10}$$

Similarly, since the function $f_{10}(z,s)$ is analytic in the domain $Re(s) \geq 0$, and since the left right hand of 7 is equal to zero for $s = \lambda - \lambda z + \theta$, then the right hand side must also be zero at this point, so we have

$$\frac{\partial Q_{10}(z,0)}{\partial x} = \Big(\lambda + \frac{\nu}{z}\Big) Q_{00}(z) h_b(\lambda - \lambda z) + \theta(\lambda - \lambda z) f_{11}(z, \lambda - \lambda z) \; . \tag{11}$$

From 10, we obtain that

$$f_{11}(z, \lambda - \lambda z) = \Big(\lambda + \frac{\nu}{z}\Big) Q_{01}(z) \frac{h_v(\lambda - \lambda z + \theta) - h_v(\lambda - \lambda z)}{(\lambda - \lambda z)\theta} \; . \tag{12}$$

So, we get

$$\frac{\partial Q_{10}(z,0)}{\partial x} = \Big(\lambda + \frac{\nu}{z}\Big) \Big(Q_{00}(z) h_b(\lambda - \lambda z) + Q_{01}(z) [h_v(\lambda - \lambda z + \theta) - h_v(\lambda - \lambda z)] \Big) \; . \tag{13}$$

Substituting now 10 and 13 in 7, we obtain the function $f_{10}(z,s)$ under the following form

$$s(s - \lambda + \lambda z) f_{10}(z,s) = \Big(\lambda + \frac{\nu}{z}\Big) \Bigg(Q_{00}(z) \big[h_b(\lambda - \lambda z) - h_b(s) \big] +$$

$$+ Q_{01}(z) \Big[h_v(\lambda - \lambda z + \theta) - h_v(\lambda - \lambda z) - \theta \frac{h_v(\lambda - \lambda z + \theta) - h_v(s)}{s(s - \lambda - \lambda z + \theta)} \Big] \Bigg) \; . \tag{14}$$

By Tauberian theorem, we have

$$Q_{11}(z,\infty) = lim_{x \to \infty} Q_{11}(z,x) = lim_{s \to 0+} s f_{11}(z,s) \; .$$

So, from equation 10, we get

$$Q_{11}(z,\infty) = \Big(\lambda + \frac{\nu}{z}\Big) \frac{1 - h_v(\lambda - \lambda z + \theta)}{\lambda - \lambda z + \theta} Q_{01}(z) \; . \tag{15}$$

Similarly, we obtain from 14

$$Q_{10}(z,\infty)=\Big(\lambda+\frac{\nu}{z}\Big)\Bigg(Q_{00}(z)\frac{1-h_b(\lambda-\lambda z}{\lambda-\lambda z}+$$

$$+Q_{01}(z)\Big(\frac{h_v(\lambda-\lambda z)-h_v(\lambda-\lambda z+\theta)}{\lambda-\lambda z}+\theta\frac{1-h_v(\lambda-\lambda z+\theta)}{(\lambda-\lambda z)(\lambda-\lambda z+\theta)}\Bigg)\ . \tag{16}$$

Now taking into account 9, the equation 4 become

$$(\lambda+\nu+\theta)Q_{01}(z)=\Big(\lambda+\frac{\nu}{z}\Big)Q_{01}(z)h_v(\lambda-\lambda z+\theta)+B\ ,$$

where

$$B=\frac{dP_{10}(0)}{dx}+\nu P_{01}(0)\ .$$

So,

$$Q_{01}(z)=\frac{B}{\lambda+\nu+\theta-\Big(\lambda+\frac{\nu}{z}\Big)h_v(\lambda-\lambda z+\theta)}\ , \tag{17}$$

where

$$R_v(\lambda-\lambda z)=\lambda+\nu+\theta-\Big(\lambda+\frac{\nu}{z}\Big)h_v(\lambda-\lambda z+\theta)\ . \tag{18}$$

Similarly, from 13 the equation 3 become

$$(\lambda+\nu)Q_{00}(z)=\big(\lambda+\frac{\nu}{z}\big)Q_{00}(z)h_b(\lambda-\lambda z)+$$

$$+\big(\lambda+\frac{\nu}{z}\big)Q_{01}(z)h_b[(\lambda-\lambda z+\theta)-h_b(\lambda-\lambda z)]+\theta Q_{01}(z)-A\ .$$

After some algebraic manipulations we can rewrite the above equation under the form

$$Q_{00}(z)=\frac{BS(\lambda-\lambda z)-AR_v(\lambda-\lambda z)}{R_b(\lambda-\lambda z)R_v(\lambda-\lambda z)}\ . \tag{19}$$

where

$$A=\theta P_{01}(0)\ , \tag{20}$$

$$R_b(\lambda-\lambda z)=\lambda+\Big(\lambda+\frac{\nu}{z}\Big)h_b(\lambda-\lambda z)\ ,$$

$$S(\lambda-\lambda z)=\theta+\Big(\lambda+\frac{\nu}{z}\Big)[h_v(\lambda-\lambda z+\theta)-h_v(\lambda-\lambda z)]\ . \tag{21}$$

So, we have proved the following

Theorem 2. *If the conditions 1-2 of theorem* 1 *are fulfilled, then the joint distribution of the server state and orbit size is given by it's generating transform 15-17-19. The constants A and B are derived in the next section.*

5 Some Performance Measures

In this section we derive some performance measures of interest.

5.1 The Probability That the Server Is Available, But on Working Vacation

If we take $z = 1$ in formula 16, we have

$$q_{01} = Q_{01}(1) = \frac{B}{\alpha} = \frac{B}{\theta + (\lambda + \nu)[1 - h_v(\theta)]} \; . \, , \tag{22}$$

where

$$\alpha = R_v(0) = \theta + (\lambda + \nu)[1 - h_v(\theta)] \; . \tag{23}$$

5.2 The Probability That the Server Is Available and Not on Working Vacation

This probability is given by $Q_{00}(1)$, but the denominator in 18 equal $R_b(0)R_v(0) = 0$. So the numerator must also be zero, and we have a relation between A and B,

$$BS(0) - AR_v(0) = 0 \; , \tag{24}$$

where

$$\alpha = R_v(0) = \theta + (\lambda + \nu)[1 - h_v(\theta)] \; . \tag{25}$$

So,

$$B = A\frac{\alpha}{\gamma} \; , \tag{26}$$

where

$$\gamma = S(0) = \theta - (\lambda + \nu)[1 - h_v(\theta)] \; . \tag{27}$$

So, we have

$$BS(0) - AR_v(0) = 0 \; . \tag{28}$$

and

$$Q_{00}(z) = A\frac{\frac{\alpha}{\gamma}S(\lambda - \lambda z) - R_v(\lambda - \lambda z)}{R_b(\lambda - \lambda z)R_v(\lambda - \lambda z)} \; . \tag{29}$$

By using l'Hospital rule, we obtain

$$q_{00} = \frac{A}{\alpha\gamma} \times$$

$$\times \, \frac{\alpha(\theta - \lambda[1 - h_v(\theta)] + (\lambda + \nu)(\tau_v^{'} - \tau_v)) - \gamma(\theta + \lambda[1 - \lambda_v(\theta)] - (\lambda + \nu)\tau_v^{'}))}{\nu - (\lambda + \nu)\tau_b} \, , \tag{30}$$

where,

$$\tau_b = -\lambda h_b^{'}(0), \qquad \tau_v = -\lambda h_v^{'}(0) \; ,$$

$$\tau_b^{'} = -\lambda h_b^{'}(\theta), \qquad \tau_v^{'} = -\lambda h_v^{'}(\theta) \ ,$$
$$\alpha\gamma = \theta^2 - (\lambda+\nu)^2[1-h_v(\theta)]^2 \ ,$$
$$M = \theta - \lambda[1-h_v(\theta)] + (\lambda+\nu)(\tau_v^{'} - \tau_v) \ ,$$
$$N = \theta + \lambda[1-\lambda_v(\theta)] - (\lambda+\nu)\tau_v^{'} \ .$$

5.3 The Probability That the Server Is Busy and on Working Vacation

Setting $z = 1$ in formula 15

$$q_{11} = \frac{A(\lambda+\nu)[1-h_v(\theta)]}{\alpha\gamma\theta} \times \frac{\alpha M - \gamma N}{\nu - (\lambda+\nu)\tau_b} \ . \tag{31}$$

5.4 The Probability That the Server Is Busy and Not on Working Vacation

Substituting 17 and 19 in 16, we obtain

$$Q_{10}(z,\infty) = \frac{A(\lambda+\frac{\nu}{z})}{\gamma R_v(\lambda-\lambda z)(\lambda-\lambda z)} \times$$
$$\times \Big([1-h_b(\lambda-\lambda z)] \frac{\alpha S(\lambda-\lambda z) - \gamma R_v(\lambda-\lambda z)}{R_b(\lambda-\lambda z)} +$$
$$+ [h_v(\lambda-\lambda z) - h_v(\lambda-\lambda z+\theta)] + \theta \frac{1-h_v(\lambda-\lambda z+\theta)}{\lambda-\lambda z+\theta} \Big) \ . \tag{32}$$

Using the l'Hospital rule two times, we obtain after some algebraic manipulations the value of P_{10} which is not provided here for constraint of space.

6 Conclusion

The obtained generating functions allow us to obtain several other performance measures such as the mean waiting or the mean sojourn time and so on. But, for lack of space it is reported for an ongoing work in which control problem of vacation policy will be investigated.

Appendix A

The unknown constant A can be obtained using the normalization condition

$$P_{00} + P_{01} + P_{10} + P_{11} = 1 \ ,$$

We found that

$$A = \frac{\gamma(\nu - (\lambda+\nu)\tau_b)}{\Phi} \ ,$$

where,

$$\Psi = (\lambda+\nu)\theta\tau_b(\alpha\delta-\gamma\xi) - 2\lambda\theta[\nu-(\lambda+\nu)\tau_b][1-h_v(\theta)] + (\lambda+\nu)\theta[2\lambda\tau_b -$$

$$-(\lambda+\nu)\lambda h_b^{"}(0)][1-h_v(\theta)] + (\lambda+\nu)[\nu-(\lambda+\nu)\tau_b](\lambda[1-h_v(\theta)]+\theta(\tau_v^{'}-\tau_v)+\theta\tau_v^{'}) \ ,$$

and

$$\Phi = \frac{\alpha(\theta-\lambda[1-h_v(\theta)]+(\lambda+\nu)(\tau_v^{'}-\tau_v)) - \gamma(\theta+\lambda[1-h_v(\theta)]-(\lambda+\nu)\tau_v^{'})}{\alpha} +$$

$$+\frac{\Psi}{\lambda\theta([\theta+(\lambda+\nu)\nu[1-h_v(\theta)])} + \frac{(\lambda+\nu)[1-h_v(\theta)][\alpha M-\gamma N]}{\alpha\theta} + \nu - (\lambda+\nu)\tau_b \ .$$

Appendix B: Heuristic Proof of Theorem 1

The conditions 1-2 of ergodicity appears from the formula 22, 30 and 31. Indeed, the quantities q_{00}, q_{01} and q_{11} are probabilities which must be strictly positive (since all the computations are valid if the corresponding Markov process is ergodic), so all the quantities appearing in the denominators of these expressions (in particular α,$\nu-(\lambda+\nu)\tau_b$ must be strictly positive, which gives the conditions 1 and 2.

References

1. Artalejo, J.R.: A Classified Bibliography on Retrial Queues: Progress in 2000-2009. Math. & Comput. Modelling 51(9), 1071–1081 (2009)
2. Do, T.V.: M/M/1 Retrial Queue with Working Vacation. Acta Informatica 47, 67–75 (2009)
3. Jain, M., Agrawal, P.K.: M/E_k/1 Queueing System with Working Vacacation. Quality Tech. & Quantit. Managmt. 4(4), 455–470 (2006)
4. Ke, J., Wu, C., Zhang, Z.G.: Recent Developments in Vacation Queueing Models-A Survey. Int. J. Oper. Res. 7(4), 3–8 (2010)
5. Li, J.H., Tian, N.S., Zhang, Z.G., Lu, H.P.: Analysis of the M/G/1 Queue with Exponential Working Vacations- A Matrix Analytic Approach. Queueing Syst. 61, 139–166 (2009)
6. Liu, W., Xu, X., Tian, N.: Some results on the $M/M/1$ Queue with Working Vacacations. Operat. Res. Letters 35(5), 595–600 (2007)
7. Servi, L.D., Finn, S.G.: M/M/1 Queues with Working Vacations (M/M/1/WV). Performance Evaluation 50, 41–52 (2002)
8. Wu, D., Takagi, H.: $M/G/1$ Queue with Multiple Working Vacacations. Performance Evaluation 63(7), 654–681 (2009)
9. Xu, X., Liu, M.X., Zhao, X.H.: The balk Input M_X/M/1 Queue with Working Vacacations. J. Syst. Sci. & Syst. Engineering 18(3), 358–368 (2009)
10. Yang, D.Y., Wang, K.H., Wu, C.H.: Optimization and Sensitivity Analysis of Controlling Arrivals in the Queueing System with Single Working Vacation. J. of Comput. & Appl. Math. 234, 545–556 (2010)
11. Zhang, M., Hou, H.: Performance Analysis of MAP/G/1 Queue with Working Vacacations and Vacation Interruption. Appl. Math. Modelling 35(4), 1551–1560 (2011)

Macroeconomic Analysis and Parametric Control Based on Computable General Equilibrium Model of the Regional Economic Union

Abdykappar Ashimov, Yuriy Borovskiy,
Nikolay Borovskiy, and Bahyt Sultanov

Kazakh National Technical University named after K. Satpayev,
22 Satpayev St., 050013, Almaty City, Kazakhstan
ashimov37@mail.ru, {yuborovskiy,nborowski86}@gmail.com,
sultanov_bt@pochta.ru
http://www.kazntu.kz/en

Abstract. The paper describes a proposed mathematical model of the regional economic union. The model relates to a class of computable general equilibrium models (CGE models). There are given results of parametric identification and verification of the model. There are also described setting and solving of parametric control problems on evaluation of economic policy tools at the level of single countries and the economic union based on a verified model. It has been shown that the problem solution for estimating optimal values of the tools at the level of the regional economic union is rational than one at the level of single countries.

Keywords: Computable general equilibrium model, Multiregional economic modeling, Model verification, Parametric control.

1 Introduction

Since 2010 there has been functioning the Customs Union (CU) of three countries (the Republic of Kazakhstan, the Russian Federation, and the Republic of Belarus), and since 2012 the Common Economic Space (CES) which unites the mentioned countries. Based on it, there is expected creation of the Eurasian Economic Union by 2015.

Implementation of this objective requires at first comprehensive vision of middle-term prospects of the interaction between countries-members of the Customs Union and the Common Economic Space and adequate tool for macroeconomic analysis and recommendations-making for optimal economic policy which considers potential effects of different external and internal factors.

There are not set any problems for estimating optimal values of economic policy tools in existing dynamic stochastic general equilibrium models [1], [2], [3] proposed for the description of the regional economic unions and in computable

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_44, © Springer International Publishing Switzerland 2014

general equilibrium models proposed for the description of effects of global and regional economic policies on ecology [4], [5], [6].

This paper is about estimation of optimal values of economic policy tools at the level of the regional economic union taking for example the Customs Union and the Common Economic Space of three countries (Kazakhstan, Russia, and Belarus). The mentioned estimation is made based on the CGE models and the theory of parametric control of macroeconomic systems.

Application of the proposed CGE model differs from existing results by the following:

- values of all its exogenous and endogenous variables economic indicators for the identification period reproduce corresponding statistical meanings, the models structure does not change in the forecasting period compared to one in the identification period;
- calculation of equilibrium values of endogenous variables in the nonlinear model is made without model linearization;
- the model describes the government sector, which incorporates an expanded interpretation of monetary and fiscal policies;
- the model describes investments into fixed assets by producers, the government, the rest of the union's countries, and the rest of the world.

2 CGE Model for the Customs Union and the Common Economic Space

The Constructed Customs Union CGE model describes a behavior and interaction of stated below economic agents of the three mentioned countries in the framework of the CU agreements as with each other, so with the rest of the world. A model of the Common Economic Space (hereinafter Model) is a CGE model of the CU with additional conditions of harmonization of macroeconomic policy in terms of three inequalities, which impose on the values of endogenous variables of the CU's model from 2012. Economic agents of the Model and their main functions are stated below (hereinafter i = 1,2,3 – serial number of the CU Country, i = 1 appropriates to Kazakhstan, i =2 – Russia, i =3 – Belarus).

Agent – Aggregate Producers (AP) of the Country i: Produce intermediate, consumer, investment products for domestic consumption, and also export products for other Countries and for the rest of the world; Consume (domestic and imported) intermediate and investment products, and also labor; Pay taxes to Government; Define demands for loans and deposits of legal entities.

Agent – Households of the Country i: Offer labor for AP of the Country i: Consume domestic and imported consumer products; Pay taxes and compulsory pension contributions to Government and receive from it subsidies; Define demands for loans and deposits of individuals.

Agent – Government of the Country i: Forms government income and government spending of the Country i; Defines government demand for domestic and imported Consumer products; Subsidizes Households and AP transfers of the Country i; Forms National fund income and National fund spending. Governments of three Countries distribute jointly collected customs duties on import among Countries.

Agent – Banks of the Country i: Define refinancing interest rate, money holding, interest rates for deposits and loans in the Country i; Meet demands for loans and deposits of AP and households of the Country i.

Agent – the Rest of the World: Define prices for export and import products to (from) the rest of the world for each Country i; fully meet demands for export and import products of Countries.

Markets of the Model are to define the prices, at which obtains corresponding equalities of demands and supplies of products (including VAT) and of labor. The Model has three markets of domestic intermediate products of each Country; three markets of domestic consumer products of each Country; three markets of domestic investment products of each Country; three labor markets of each Country; six markets of export (import) products for each pair of Countries;

General view of the Model is presented by the following system [7], [8].

1. Subsystem of difference equations, linking values of variables $x_1(t)$ (outputs, fixed assets of agents-producers, account balances of agents in banks and others for the three above mentioned countries) for two successive years:

$$x_1(t+1) = f_1(x_1(t), x_2(t), x_3(t), u(t), a(t)) \ , \ x_1(0) = x_{1,0} \ . \qquad (1)$$

Here $t = 0, 1, \ldots, (n-1)$ – serial number of year, discrete time; $x(t) = (x_1(t), x_2(t), x_3(t)) \in R^m$ – vector of all endogenous variables of system, describing statuses of economies in the three countries of economic union;

$$x_i(t) \in X_i(t) \subset R^{m_i} \ , \ i = 1, 2, 3 \ . \qquad (2)$$

Here $m_1 + m_2 + m_3 = m$; $x_2(t)$ – demand and supply values of agents in all markets and others; $x_3(t)$ – different types of market prices; $u(t) \in U(t) \subset R^q$ – vector function of controllable parameters. Coordinate values of this vector correspond to different government economic policy tools of mentioned three countries, for example, as different tax rates, refinancing interest rates, money holdings, etc.; $a(t) \in A \subset R^s$ – vector function of uncontrollable parameters (factors). Coordinate values of this vector describe different depending on time external and internal social and economic factors of union's countries: prices for different kind of export and import productions, labor quantity, production function parameters, etc.; $X_1(t)$, $X_2(t)$, $X_3(t)$, $U(t)$ – compact sets with non-empty interiors; $X_i \in \cup_{t=1}^{n} X_i(t)$, $i = 1, 2, 3$; $X \in \cup_{i=1}^{3} X_i$; $U \in \cup_{t=0}^{n-1} U(t)$; A – open connected set; $f_1 : X \times U \times A \to R^{m_1}$ – continuous mapping.

2. Subsystem of algebraic equations, describing behavior and interaction of agents in different markets during chosen year, these equations assume expressing variables $x_2(t)$ via rest endogenous variables for chosen exogenous functions $u(t)$ and $a(t)$:

$$x_2(t) = f_2(x_1(t), x_3(t), u(t), a(t)) \ , \tag{3}$$

$f_2 : X_1 \times X_3 \times U \times A \to R^{m_2}$ – continuous mapping.

3. Subsystem of recurrence relations for iterative solution of market prices' equilibrium values in all markets of the Model:

$$x_3(t)[Q+1] = f_3(x_2(t)[Q], x_3(t)[Q], u(t), a(t), L) \ . \tag{4}$$

Here $Q = 0, 1, \ldots$ – serial number of iteration; L – set of positive numbers (adjustable iteration constants; economic system faster obtains its equilibrium as their values decrease, however the risk of price shifting to the negative side increases simultaneously; $f_3 : X_2 \times X_3 \times U \times A \times (0, +\infty)^{m_3} \to R^{m_3}$ – continuous mapping (joint with f_2) is contracting at fixed t, $x_1(t) \in X_1(t)$ and some fixed L. In this mapping case the mappings f_2, f_3 have the only fixed point, to which leads the iterative process (3), (4).

CGE model (1), (3), (4) at fixed values of exogenous functions $u(t)$ and $a(t)$ for each moment t defines values of endogenous variables $x(t)$, appropriate to demand and supply equilibrium prices in all markets of the Model.

3 Parametric Identification and Verification of the Model

Parametric identification (calibration) of the Model has been performed in three stages.

On the first stage, the parameters of multiplicative production functions which determine the values of gross outputs by the aggregate producers of all CU Countries depending on factors of production (fixed assets, labor, intermediate products, and imported oil) were evaluated.

On the second stage, the values of exogenous functions $u(t)$, $a(t)$ of the Model for the historical period (2000–2011) were taken based on observed statistical data of the Countries and the rest of the world.

On the third stage, the values of correcting coefficients from the Model's corresponding equations for the period 2000–2011 were determined based on observed statistical data for exogenous and endogenous variables of the Model.

The evaluated model accurately reproduces the statistical data of 362 endogenous variables of the Model for the period 2000–2011. Basic calculation of the Model for the period 2000–2018 is made by forecasting exogenous functions and coefficients of the Model for 2012–2018.

Verification of the evaluated Model has been made through estimation of stability indicators, retroforecast and estimation of sensitivity coefficients.

The stability indicator of the Model is a diameter of ball's (with the one percent radius and the center at the point of some exogenous parameters of the Model) image in relative values for (set by the Model) mapping from exogenous variables onto endogenous ones. Here, as exogenous parameters were taken various types of external prices, output shares and expense shares of all three countries' AP, and others for 2000. As output parameters, there were taken GDP and CPI of the CU countries for the chosen year. All obtained stability indicators' results do not exceed 9.93, which characterizes the stability of the Model when estimated till 2018 as sufficiently high (see the Table 1).

Table 1. Stability indicators of the Model

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Indicator	0.96	1.54	2.11	2.54	1.69	3.31	4.01	4.46	5.25	5.34
Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Indicator	5.85	6.58	7.52	6.94	7.98	8.08	8.66	9.19	9.93	

The verification of the Model with **retroforecast** is made as following.

- With observed data for 2000–2010, there was created a version of the Model.
- Corresponding values of all endogenous macroeconomic indicators of the Model were calculated based on extrapolation of exogenous variables of the Model version for 2011–2012.
- The relative root-mean-square deviation of all calculated values for 2011–2012 from corresponding observed values was about 2.9 percent.

The verification of the Model was also made through estimation of **sensitivity (elasticity) coefficients** for values of endogenous variables of the Model by its exogenous parameters to verify the compliance of signs of obtained estimates with main tenets of the macroeconomic theory. The following Table 2 shows estimation of sensitivity coefficients for two variables of Kazakhstan – GDP and Consumer price level (CPL) calculated on the basis of the Model.

The verification results of the Model by three approaches show the Models acceptable adequacy.

4 Macroeconomic Analysis of Causal Factors of the 2009 Recession Based on the Model

One of the directions of macroeconomic analysis based on the Model aimed to determine reasons of macroeconomic events which were related to basic macroeconomic indicators of the Countries changing during the crisis period in 2009.

In the framework of solving this problem, there was evaluated the sensitivity of impact of the following parameters (external and internal exogenous factors including government policy tools for 2008–2009) for assigning GDP variables (Yg_i) and the consumer price index (P_i) for 2009:

1. prices on the Countries export products into the rest of the world (Pex_i);
2. prices on various products imported to the Countries from the rest of the world (PcI_i, PzI_i, PnI_i);

Table 2. Sensitivity coefficients

Parameter (2008)	GDP (2009)	CPL (2009)
Price of non-oil export products	0.23	1.24
Price of imported consumer products from the rest of the world	-0.06	-0.94
Price of imported intermediate products from the rest of the world	0.00	-0.89
Price of imported investment products from the rest of the world	-0.06	-0.62
Technological coefficient of gross output	1.03	0.64
Intermediate products' share in output	0.03	0.02
Consumer products' share in output	0.00	0.07
Investment products' share in output	0.00	-0.01
Export products' share in output	0.02	0.01
Consumption share of AP intermediate products	0.00	0.04
Consumption share of AP investment products	-0.01	-0.01
Share government spending in the state budget	0.21	0.41
Effective rate of CIT (corporate income tax)	-0.37	0.29
Refinancing rate shock	-0.18	-0.40
Money holding shock	0.07	0.12
Oil price	0.26	1.26

3. technological coefficients of the gross output production functions of the Countries (Y_i);
4. the share of AP production of various products in the Countries (Ez_i, Ec_i, En_i, Eex_i);
5. the share of AP consumption of various products in the Countries (Oz_i, On_i);
6. the share of government consumption in government spending of the Countries (G_i);
7. effective rates of corporate income tax of the Countires (T_i);
8. refinancing rates of the Countries (Ref_i);
9. money holdings of the Countries (DB_i);
10. oil price ($Poil$).

Analysis of calculated elasticity coefficients (Table 2) shows that the impact of output shares (Ec_i, En_i, Eex_i) and consumption shares (Oz_i, On_i) on studied macroeconomic indicators is little enough.

Further, by using counterfactual scenario analysis the impact degree of indicated above parameters on variables Yg_i and P_i fluctuations was evaluated in accordance with the following algorithm.

1. Ten scenarios are calculated in which one j-parameter of the given list (except Ec_i, En_i, Eex_i, Oz_i, On_i) remains in 2008 and 2009 at the level of one in 2007, and the rest of indicators from the list are statistical. Corresponding increments of the variables Yg_i and P_i compared to baseline values are obtained: ΔYg_{ij} and ΔP_{ij}.
2. The scenario in which all mentioned ten parameters in 2008 and 2009 remain at the level of those in 2007 is calculated. Corresponding increments of the variables Yg_i and P_i compared to baseline values are ΔYg_i and ΔP_i.
3. Relationships of $\Delta Yg_{ij}/\Delta Yg_i$ and $\Delta P_{ij}/\Delta P_i$ (in %) are calculated. These relationships characterize the impact degree of corresponding factors on the increments of indicators.

It is worth to note that if the values of mentioned parameters for 2008–2009 remained at the level of those in 2007 the real GDP of Kazakhstan for 2009 would have been higher than the observed one by 11.8%, and CPI by 3.7%. The results of mentioned impact degrees for Kazakhstan are presented in the Table 3.

Table 3. Impact degrees of parameters fluctuations on increments of the variables in 2009 (in %)

	Parameter											
Variable	Pex_1	PcI_1	PzI_1	PnI_1	Y_1	G_1	T_1	Ref_1	DB_1	$Poil$	Others	Total
Yg_1	-62.0	1.2	0.8	1.2	-64.0	61.2	5.5	30.1	36.1	-88.5	-18.0	-100
P_1	-53.7	3.2	2.4	2.1	-6.5	19.8	0.7	11.0	10.0	-70.7	-16.6	-100

Analysis of the Table 3 shows that state policy measures in 2008–2009 were correct but not optimal (as the following results of parametric control indicate).

5 Parametric Control Problems of the Regional Economic Union Based on the Model

Next group of experiments on parametric control problems solving [7] was made during evaluation of counterfactual optimal values of budgetary and fiscal policy tools of the CU Countries for 2007–2011 in case of absence and presence of coordination such policies. Here are four such Problem Pr_i $(i = 0, 1, 2, 3)$ informal definitions, where the values of uncontrollable exogenous variables of the Model correspond with basic (retrospective) prognosis of these variables.

Pr_i **Parametric Control Problems Setting.** On the basis of the Model, to find the values of tax rates and shares of government spendings in budgets for 2007–2011 for each Problem Pr_i, which provide maximum of criterion K_i, $(i = 0, 1, 2, 3)$ at corresponding restrictions for controllable parameters and some endogenous variables to meet the conditions of debt stability and competitiveness of the CU Countries. Here $i = 1, 2, 3$ – serial number of the CU Country, criterion K_i – average real GDP of the Country i for 2013–2017, only government policy tools of the Country i are used. In the Problem Pr_0 criterion K_0 is average real total GDP of three CU Countries for 2013–2017, and applying government policy tools consist of corresponding tools of three CU Countries.

Increments of the mentioned criteria K_i (in percent relative to basic variant), corresponding with computational solutions for Problems Pr_i are illustrated in the Table 4, and the CU GDP diagrams are on the Fig. 1. An analysis of the Table 4 indicates that in the framework of Problems Pr_i $(i = 0, 1, 2, 3)$ an approach of parametric control on the level of all Union Countries provides effect for each Union Country not less (for two Countries larger) than parametric control on the level of single Country.

Table 4. Four parametric control problems solutions results

	Increment of criterion (in %)			
Problem	K_1	K_2	K_3	K_0
Pr_1	4.05	0.64	0.14	0.58
Pr_2	0.78	3.68	1.75	2.36
Pr_3	0.25	0.43	3.83	0.32
Pr_0	4.07	3.68	4.06	3.77

To make recommendations on optimal state policy of the CU countries the following parametric control problem has been solved for 2014–2018.

Setting the Problem 2. Based on the Model, to determine values of economic tools (effective rate of CIT, share of government consumption in government spending) at the level of each CU country for 2014–2018 which allow the maximum of Kr criterion at corresponding limits of the coordination Indicators of macroeconomic policies and of the values of these economic tools.

Here: $Kr = a1 \times K1 - a2 \times K2 + a3 \times K3 - a4 \times K4 - a5 \times K5$;

$K1$ – Normalized average (for 5 years) value of the CUs GDP per capita, in USD;

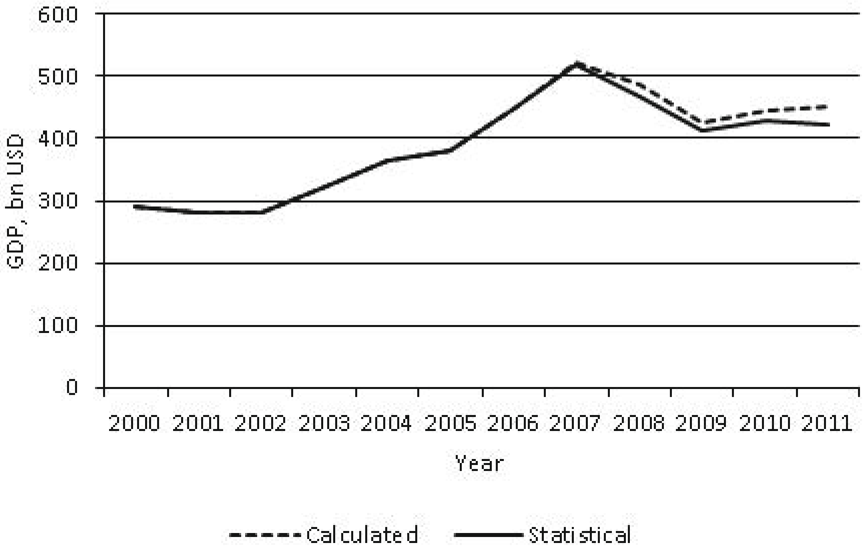


Fig. 1. CU GDP in bn USD, in prices of 2000

$K2$ – Normalized average (for 5 years) value of the government debt in the CU countries, in million USD;

$K3$ – Normalized average (for 5 years) value of export from the CU countries, in million USD;

$K4$ – Normalized average (for 5 years) value of import into the CU countries, in million USD;

$K5$ – Normalized criterion which characterizes the convergence of the CU countries by rates of GDP, CPI and the ratio of the government budget deficit to GDP;

aj $(j = 1, \ldots, 5)$ – weight coefficients, in the example $aj \equiv 1$.

The following Tables 5 and 6 illustrate the results of Problem 2 solving by numerical procedure of the Nelder-Mead algorithm. Here, Kji are components of the criterion Kj $(j = 1, \ldots, 4)$, pertaining to the Country i $(i = 1, 2, 3)$.

Table 5. Kj criteria changes relative to basic variant (in %)

Criterion	$K1$	$K2$	$K3$	$K4$	$K5$
Change	+3.71	-3.87	+6.61	-3.22	-2.88

Table 6. *Kij* criteria changes relative to basic variant (in %)

	Criterion			
Country	$K1i$	$K2i$	$K3i$	$K4i$
Kazakhstan ($i = 1$)	+3.78	-3.71	+6.63	-3.36
Russia ($i = 2$)	+3.65	-3.84	+6.57	-3.20
Belarus ($i = 3$)	+3.71	-4.21	+6.40	-3.69

Analysis of the Tables 5 and 6 shows high potentials of the parametric control approach for making recommendations on coordinated optimal state economic policies of the regional economic Union's Countries.

6 Conclusion

1. A computable general equilibrium model for the regional economic union has been proposed taking for example the Customs Union.
2. Effectiveness of parametric control theorys application for estimation of optimal values of economic policy tools has been shown.
3. Preference for solution of the estimation problems of values of economic tools at the level of the regional economic union rather than at the level of single countries of the union has been illustrated.
4. Obtained results could be used for solving practical problems in economic policies of regional economic unions.

References

1. Coenen, G., McAdam, P., Straub, R.: Tax Reform and Labour-Market Performance in the Euro Area. A Simulation-based Analysis using the New Area-wide Model. ECB Working Paper Series 747 (2007)
2. Brubakk, L., Husebo, T.A., Maih, J., Olsen, K., Ostnor, M.: Finding NEMO: Documentation of the Norwegian Economy Model. Norges Bank Staff Memo 6 (2006)
3. Senaj, M., Vyskrabka, M., Zeman, J.: MUSE: Monetary Union and Slovak Economy Model. NBS Working Paper 1 (2010)
4. Babiker, M.H., Reilly, J.M., Mayer, M., Eckaus, R.S., Wing, I.S., Hyman, R.C.: The MIT Emissions Prediction and Policy Analysis (EPPA) Model: Revisions, Sensitivities, and Comparisons of Results. Report No. 71 (2001)
5. Current GTAP Model, `https://www.gtap.agecon.purdue.edu/models/current.asp`
6. MIRAGE Model, `http://www.mirage-model.eu`
7. Ashimov, A.A., Sultanov, B.T., Borovskiy, Y.V., Adilov, Z.M., Novikov, D.A., Alshanov, R.A., Ashimov, A.A.: Macroeconomic analysis and parametrical control of a national economy, p. 288. Springer, New York (2013)
8. Makarov, V.L., Bakhtizin, A.R., Sulakshin, S.S.: The use of computable models in public administration, p. 304. Scientific Expert, Moscow (2007) (in Russian)

Self-organizational Aspects and Adaptation of Agent-Based Simulation Based on Economic Principles

Petr Tučník, Pavel Čech, and Vladimír Bureš

Department of Information Technologies, Faculty of Informatics and Management,
University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
{petr.tucnik,pavel.cech,vladimir.bures}@uhk.cz

Abstract. The agent-oriented approach is one of most frequently used techniques for complex system simulation today. This paper is investigating application of multi-agent system consisting of four basic types of agents for creating virtual economy environment for further testing and research in areas of multi-agent coordination and self-organization. Although the proposed system is in several aspects simplified, for example banking sector and government are not included into model, it provides useful basis for research of adaptation mechanisms, manufacturing management, supply chain management, and customer behaviour modelling. Individual goals and strategies are forming collective effort of pursue of given goals, respecting constraints and limitations set on level of the whole agent community. Our goal is to design a system consisting of agents capable of self-organization into structures allowing processing of resources in the given environment and creating production and supply chains with maximum efficiency possible.

Keywords: Agent, modelling, simulation, virtual economy, self-organization, adaptation.

1 Introduction

Agent-based modelling is a quite popular modelling approach which is widely used in many disciplines. Current scientific papers indexed in the Web of Science database are classified to overall 138 research areas. Whereas the majority of research papers belong to the field of computer science or engineering (67,6 %); plethora of other application areas such as toxicology, entomology, oceanography, crystallography, or management of biological incidents [1] can be identified.

In the business and economics realm the agent-based modelling is applied from the earliest stages of development of this programming paradigm. For instance, Janssen and Vries [2] develop a simple dynamic model of the economy-energy-climate system and prove that the adaptive behaviour can be included in global change modelling. Vidal and Durfee [3] used an economic multi-agent system to determine when agent should behave strategically (i.e. learn and use models of other agents), and when it should act as a simple price-taker. Their results show how savvy buyers can avoid being cheated by sellers, how price volatility can be used to quantitatively predict the

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_45, © Springer International Publishing Switzerland 2014

benefits of deeper models, and how specific types of agent populations influence system behaviour. Trading between buyers and sellers represents a segment of economics which was often selected for application of multi-agent modelling and simulation principles. Systems such as Kasbah [4] or Magma [5] can serve as examples.

Since the first pioneering studies were published the application area in the business and economics field has extended. Guessoum et al. [6] propose a new adaptive multi-agent model that includes the organizational forms into the economic models. Babita, Rao and Shukla [7] investigated possibilities of multi-agent systems in the e-business realm. Damaceanu and Capraru [8] focus their attention on banking market. In their study, they conduct 11 computer experiments and study the evolution of various banking market indicators such as total amount of money, savings, wallets, or bank reserves. Due to its complexity, Sinha et al. [9] studied and created model of petroleum supply chain. Dosi et al. [10] develop an evolutionary model of output and investment dynamics yielding endogenous business cycles. The model describes an economy composed of firms and consumers. Whereas firms belong to two industries, consumers sell their labour and consume their income. Simulation results show that the model is able to deliver self-sustaining patterns of growth characterized by the presence of endogenous business cycles.

2 Problem Formulation

Although the broader focus of application domains and shift to more complex models is apparent, the literature review reveals that several main areas can be identified in the business and economics modelling. However, these areas are mostly independent from the research perspective and do not allow simulation and analysis of mutual interrelationship that represents typical feature of the real economic systems. Therefore, endeavour should be devoted to mutual connection of the following economic segments and related issues, which individually represent a challenge from the informatics perspective:

1. Logistics – path finding and application of the graph theory; mechanisms of group transport coordination; dynamical route changes.
2. Consumer behaviour – accordance with existing theories in economics; content of the consumer basket and its determinants such as taxing, education, or social level.
3. Production processes (manufacturing) – production chain management; standardisation; achieving a certain quality level with existing technological limitations.
4. Supplying processes – supply chain management; continuity of processes with minimisation of delays; relationship of volume to number of transportation agents.
5. Managerial decision-making and planning – level of autonomy; pricing; decisions related to organisational development.
6. Labour market – education and qualification issues; accordance with existing theories; structural differentiation in an economy.

7. Services – composition of services; influence of consumer utility function; (dis)similarity to tangible products.
8. Representation of environment – maps utilisation; infrastructure; mobility of agents.

3 Model Description

Virtual economy presented in this paper represents the production and consumption processes in simulated economy, although based on real data to reasonable extent (provided by Czech Statistical Office, see also http://www.czso.cz). The aim is to simulate economic principles of effective price and quantity setting under specific demand and capacity constraints [11]. Hence, the focus is on trading products and services and offering work on a labour market. Virtual economy simulation is similar to the work of Deguchi et al. [12], however, in that representation the entities considered are more specific producing more complicated net of relations than necessary. Similarly, trust issues as discussed for example in [13] and similar concerns are not of primary attention in the presented virtual economy.

The studied virtual economy consists of four types of agents: a) consumer, b) factory, c) mining and d) transport. Due to the simplicity and clarity of relations and transparency of design, other sectors such as the banking sector or government are not included in the model. Moreover, the model represents two-sector closed economy. The basic architecture of the virtual economy depicting entities and their relations is given in Figure 1.

Consumer agent embodies the economic entity that consumes products and services (i.e. goods) and offers work. Consumer agents can buy goods based on the wealth they possess. The wealth of a consumer agent is a product of work and qualification (the higher qualification the faster accumulation of wealth). A consumer agent makes a trade-off between investment into higher qualification (e_c) and consumption. The combination of products consumed and the speed of consumption is given by the consumption function. The combination of products forms a pattern of consumption that can be used to divide consumers into three categories. The three categories are low income, middle income and high income consumers. The pattern determines the ratio of goods that the consumer agent is buying. There three types of goods: necessity, normal, luxurious. For example the proportion of goods bought by a low income class consumer might be 70 % of necessity goods such as food and household services; 20 % of normal goods and 10 % of luxury goods. The willingness to buy a certain product depends on the stock. The lower the stock of that particular product the higher price is consumer agent willing to pay (see Figure 2). In other words, the scarcity increases acceptable price. This principle is corresponding with standard price and demand relationship. The price p_{max} is the amount a consumer is willing to pay when the stock of that product is empty. Conversely, as the stock is close to 100% of the capacity the price approaches zero i.e. the consumer is willing to buy only if the price is very low.

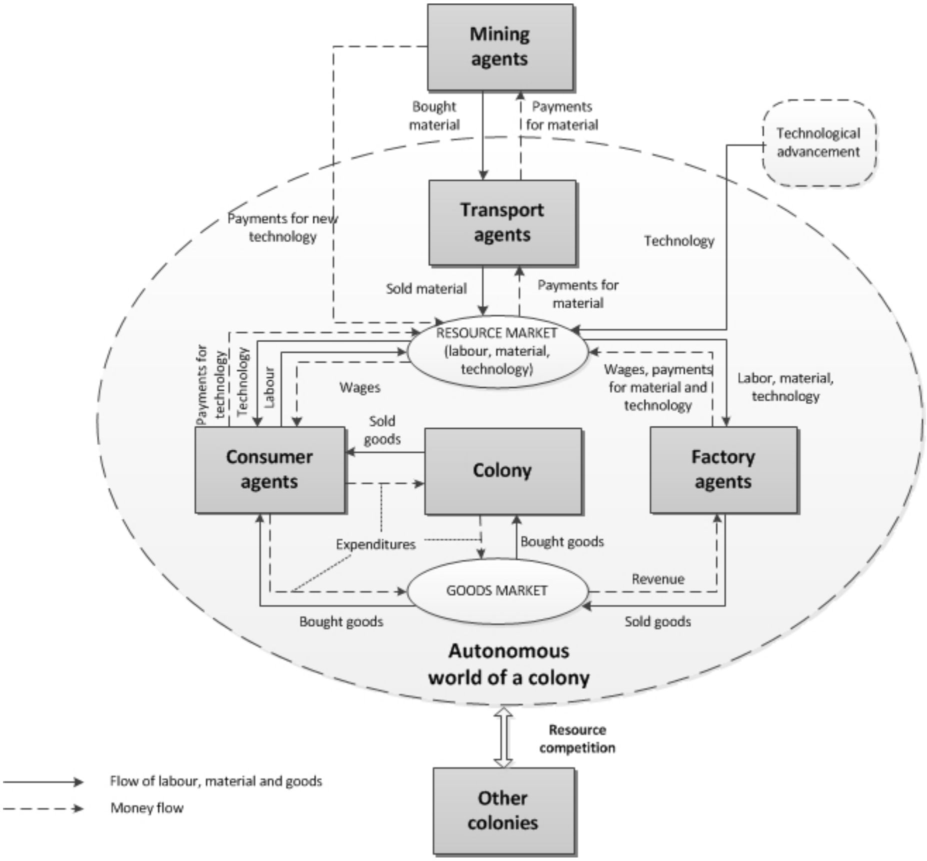


Fig. 1. Main components of model of virtual economy

The second type of agent, a factory agent, corresponds with a company in a real economy. Factory agent is responsible for transforming input to output i.e. material and other products to final product that is bought by consumer agent or sub-product that is used by another factory agent. The consumption function determines materials and their proportions. The production function determines the portfolio of goods produced. Production requires workforce i.e. employing consumer agents. The production depends on the technological level e_f and qualification of the workforce i.e. employed consumer agents e_c. The production equation is as follows:

$$\sum_{i=1}^{n} k_i^{con} x_i + WF \xRightarrow{production} e_C e_F \left(\sum_{j=1}^{m} k_j^{pro} y_j \right) \tag{1}$$

Let k_i^{con} be the speed of consumption of a material x_i and WF is the workforce; e_c is qualification level of a consumer agent and e_f technological level factory agent; k_j^{pro} be the speed of production of a product y_j.

Third type of agent in the model is mining agent. This is agent responsible for transforming resources, located in the environment, into raw material that is used by factory agents in production of goods. The cost of mining is determined by the consumption function in which the energy and technology necessary for mining is reflected. The function is similar to consumption function of a factory agent. Each mining agent supplies only one type of raw material (if several types of raw materials are produced simultaneously, each is represented by single specialized agent). Raw material, as transformed from resources, is stocked in order to be later sold to transport agents and distributed throughout the processing facilities (i.e. factories).

Transport agents serve as intermediary between mining agents and factory agents. The cost of transportation is given by the distance. There might be barriers on the way from mining agent to the factory agent; hence, it is the task of the transportation agent to find a route that is the most economical or otherwise efficient. Different strategies may be used for solving path-finding and distribution problems, e.g. transport agents may co-ordinate transportation effort with each other in order to achieve maximum efficiency. The performance of a transportation agent is determined by the speed (or mobility), capacity and technological level. Transport agent is a proxy for a particular factory agent. Thus, transport agent does not have any wealth and is buying material on behalf of a factory agent. The technological advancement of a transportation agent is also the same as for the factory agent. Transportation agent is always buying all available material up to the capacity of transportation. Transported material that is not used directly in production is stored in factory agent`s warehouse.

The modelled virtual economy contains also representation of a society of agents which is called “colony” in this context in order to avoid confusion with possible sociological semantics. The colony consists of consumer, factory and transportation agents. Mine agents do not belong to any colony, as they are distributed throughout the environment, depending on the resources they process. The colony is characterized by its position in the environment and size of population. Colonies compete for resources that are supposed to be scarce. Colony exists in an environment. The environment is important in respect to transportation provided by transportation agents.

The success of a colony can be measured by several factors. The most common efficiency metric is wealth. The wealth of a colony is given by the sum of wealth of all agents. Due to different colony populations the comparison among colonies requires computing wealth per agent. The formula is as follows:

$$cw_{COL} = \frac{\sum_{i=1}^{n}\left(w_{C,i} + w_{F,i}\right) + w_{COL}}{p} \tag{2}$$

where

$$p = \sum_{i=1}^{n} c_{COL,i} \tag{3}$$

The model enables for various configuration and thus for conducting specific experiments. These experiments might be focused on thriving of big colonies with a large number of a consumer agent. Other experiment might include the competition among colonies in case of a universal resource that cannot be substituted in the production process. Similarly, an interesting experiment might reveal the speed of wealth accumulation in case of one large colony as compared to a number of smaller competing colonies.

4 Model Experiments

4.1 The Model Purpose

The effort is focused on research of application of different strategies and effectiveness of agent`s communities in different settings. Our aim is to investigate self-organizing capabilities of the system using models of behaviour from economics theory for modelling behaviour of individual agents.

It is necessary to notice that the proposed system is not primarily focused on providing full-scale simulation of the real-world economy. The economy background of the model is rather used as a foundation for study and research in areas of self-organization and maximization of production potential of the system where population size, resources and competition are effectively defining problem domain. The goal here is to focus on agent`s decision making and control on both individual and community level. Different strategies, priorities and work distribution approaches may be studied and compared with each other within this model, making it very useful platform for research of agent`s behaviour, collaboration and co-ordination.

The organized system must have following attributes:

- Stability – system should be able to maintain itself for as long as possible, ideally for unlimited time.
- Efficiency – system is optimizing resource allocation according to its own (given) production capabilities.
- Effective distribution – system should provide capacity for suitable distribution of products. Shortages and scarcity should be avoided.
- Adaptation – system should be able to re-allocate resources and re-organize its structure according to changing conditions in the environment.

The economical context in which proposed system operates is used to maximize operational potential of the whole multi-agent system. Economic notions like "customer satisfaction" or "maximization of profit" are excellent to define target parameters for performance of individual entities in the system. These autonomous entities may then adopt different strategies in order to achieve appropriate level of efficiency in their actions. It is our intent to achieve desired behaviour of the whole system by behavioural patterns that will emerge by interaction of large amount of small individual agents with each other.

The work is divided into two phases. In the first phase a case of a single multi-agent community (colony) is investigated and researched. This phase is focused on efficiency and productiveness of the colony and its self-organizing capabilities.

The second phase of work is focused on multiple colonies interacting with each other within the given environment. It is a basic assumption that there will be several (or all) resources in the environment present in a limited amount only. In this case, individual colonies have to negotiate distribution or ownership of such resources. It is assumed that colonies would be encouraged to specialize in their production, according to given allocation of resources in the environment. This will lead to increase of mutual dependency of colonies on each other and emphasize need for their efficient cooperation. This creates excellent basis for further research in areas of agent coordination and cooperation. Application of scenarios focused on maximizing performance of colonies against each other is planned in the final phase of the project.

Table 1. Problem areas in two main branches of research

Single colony case
Individual agent behaviour
Control & strategies
Logistics & distribution
Manufacturing
Workforce allocation

Multiple colonies case
Competition issues
Resource sharing
Trade between colonies
Negotiation
Outsourcing (result of specialization)

Tab. 1 shows main areas of interest in both branches of research: single colony and multiple colonies. Take into consideration that single colony case does not mean strictly individual agents – it is also focused on organization of agent society within one community and within production units as well. Since it is expected that colonies will converge to specialization in areas of resource processing and production, colonies may be forced to work together (collaborate) in spite of the fact they are competing for resources at the same time. This increases need for efficient negotiation algorithms and strategic planning at the level of the whole agent colonies.

4.2 Example Scenario

In order to clarify the presented ideas and model, a short, simplified, example of model scenario will be presented in this part of the text. This example scenario was created in prototype version of software, implemented in NetLogo environment[1]. In this scenario, there is an M-agent producing one type of material needed at 4 colonies of different size. Fig. 2 shows user interface used in prototype application.

[1] Software platform for agent-oriented modelling. See also
http://ccl.northwestern.edu/netlogo/
for more information about NetLogo.

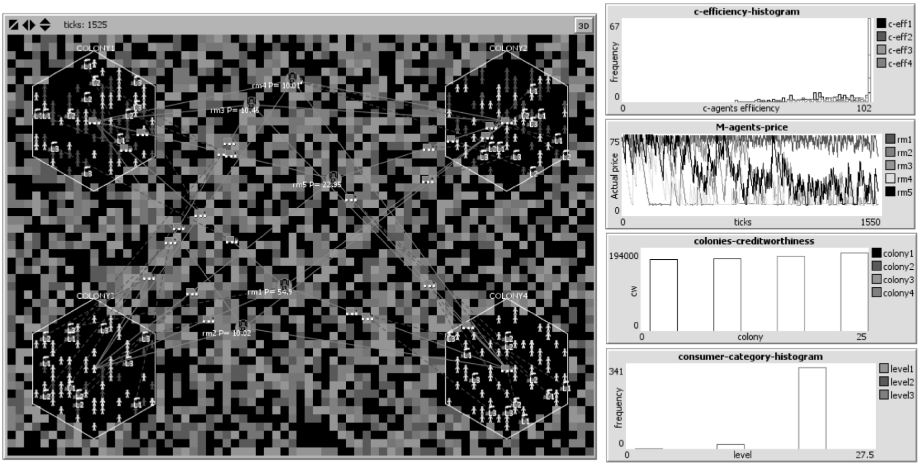


Fig. 2. User interface of prototype application, NetLogo version

In the Fig. 2, there is a graphical representation of the model environment at the left side of the screen. On the right side, there are measured attributes of the model such as satisfaction of population in individual colonies, price of product, etc. The price chart of produced material is shown at the Fig. 3.

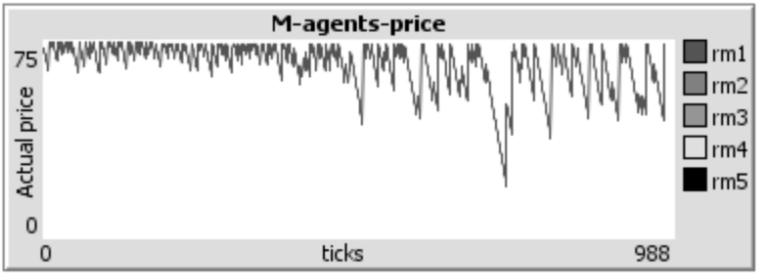


Fig. 3. Development of price of material "rm1" in 1000 iterations

The Fig. 3 shows that during first 500 steps, the M-agent is not able to produce material fast enough to satisfy users` demand, therefore, the price holds at high level. Since this situation is very favourable for M-agent, it is able to cumulate wealth quickly and after 500 steps purchase technological upgrades and increase level of production to satisfy more customers. Because individual colonies are well saturated now, the price of material drops a little occasionally (price they are willing to pay is derived from size of reserve). At this point, M-agent strategy should be adjusted to maximize its profits, becoming efficiently a price-maker for the given commodity.

From the perspective of the colony, the satisfaction level is heavily dependent on the size of the colony. In this case, 4 colonies are considered (their data are labelled c-eff1, c-eff1, etc.) with size of 100, 75, 50 and 25 agents respectively. Situation is shown at the Fig. 4. Best results are achieved by colony nr. 4, due to its small size (it is easier to satisfy demand of smaller community of agents) where satisfaction has quickly risen to high values. Colony nr 1 was able to fully satisfy demand of its inhabitants for first 200 steps only (probably also because of some small supply of material in reserve given at the initialization of the model). This resulted in unsatisfactory saturation of C-agent population inside of the colony (which reached even critical levels).

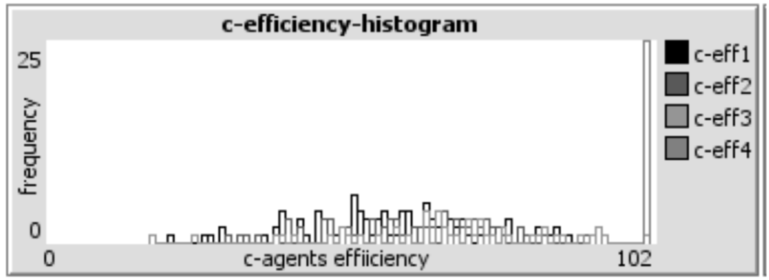


Fig. 4. Satisfaction of inhabitants (C-agents) of 4 colonies with the size of population of 100, 75, 50 and 25 agents (respectively)

This concise example was used only to demonstrate possible ways of work and decision making situations with the proposed model. It does not represent results of full scale research. More elaborate and complex scenarios can be arranged to study emergent behaviour of the agent collectives, evaluating different strategies and approaches. Idea is to fully utilize the agents as autonomous entities able to act both independently and in collaboration with each other while pursuing their individual goals.

5 Conclusions

The literature review shows that there is a strong tendency for formal defining of organizational structures and policies in self-organizing systems. Also, the agent-oriented approach is very frequently used. All this is coherent with pointing of proposed project.

The proposed model is intended to be variable, modular and adaptable. Consisting of individual agents of transparent architecture, the complex behaviour emerges over time as a result of their mutual interactions. Individual agents are pursuing their goals of maximizing profit or satisfaction (utility) according to traditional economical models. By defining consumption patterns for consumer agents, system output can be modified to produce selected type of goods or services. This will allow us to study and investigate wide range of scenarios under changing conditions in the future research.

The experiment scenario shows that in situation with scare resources it is more efficient to form smaller colonies since it is easier to cover the demand and achieve higher levels of satisfaction. The scarcity of resources also distorts the market since the resource owners are put in to the position of price makers and hence creating the conditions for imperfect competition.

Acknowledgements. Authors would like to kindly thank Miroslav Brabenec for his help with the model development and implementation. This paper is created with the support of the excellence project "Agent-based models and Social Simulation".

References

1. Bureš, V., Otčenášková, T., Čech, P., Antoš, K.: A Proposal for a Computer-Based Framework of Support for Public Health in the Management of Biological Incidents: the Czech Republic Experience. Perspect. Public Heal. 132(6), 292–298 (2012)
2. Janssen, M., de Vries, B.: The battle of perspectives: a multi-agent model with adaptive responses to climate change. Ecol. Econ. 26(1), 43–65 (1998)
3. Vidal, J.M., Durfee, E.H.: Learning nested agent models in an information economy. J. Exp. Theor. Artif. Int. 10(3), 291–308 (1998)
4. Chavez, A., Maes, P.: Kasbah: An Agent Marketplace for Buying and Selling Goods. In: 1st International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology, pp. 75–90. Practical Application Co Ltd., London (1996)
5. Tsvetovatyy, M., Gini, M., Mobasher, B., Wieckowski, Z.: MAGMA: An Agent-Based Virtual Market for Electronic Commerce. Appl. Artif. Intell. 11(6), 501–523 (1997)
6. Guessoum, Z., Rejeb, L., Durand, R.: Using adaptive multi-agent systems to simulate economic models. In: 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 68–75. IEEE Computer Society, Washington (2004)
7. Babita, M.J., Rao, M.V.G., Shukla, P.: An Agent Based Architecture for E-Business Application with Multi Agent Systems. Int. J. Adv. Eng. Appl. 3, 205–209 (2011)
8. Damaceanu, R.C., Capraru, B.S.: Implementation of a Multi-Agent Computational Model of Retail Banking Market Using Netlogo. Metal. Int. 17(5), 230–236 (2012)
9. Sinha, A.K., Aditya, H.K., Tiwari, M.K., Chan, F.T.S.: Agent oriented petroleum supply chain coordination: Co-evolutionary Particle Swarm Optimization based approach. Expert Syst. Appl. 38(5), 6132–6145 (2011)
10. Dosi, G., Fagiolo, G., Roventini, A.: The microfoundations of business cycles: an evolutionary, multi-agent model. J. Evol. Econ. 18(3-4), 413–432 (2008)
11. Pennings, E.: Price or quantity setting under uncertain demand and capacity constraints: An examination of the profits. J. Econ. 74(2), 157–171 (2001)
12. Deguchi, H., Terano, T., Kurumatani, K., Yuzawa, T., Hashimoto, S., Matsui, H., Sashima, A., Kaneda, T.: 24. Virtual Economy Simulation and Gaming –An Agent Based Approach. In: Terano, T., Nishida, T., Namatame, A., Tsumoto, S., Ohsawa, Y., Washio, T. (eds.) JSAI-WS 2001 Workshops. LNCS (LNAI), vol. 2253, pp. 218–226. Springer, Heidelberg (2001)
13. Gazda, V., Gróf, M., Horváth, J., Kubák, M., Rosival, T.: Agent based model of a simple economy. J. Econ. Interact. Coor. 7(2), 209–221 (2012)

Task-Based Modelling of the Triage Domain Knowledge

Muthukkaruppan Annamalai[1], Shamimi A. Halim[1],
Rashidi Ahmad[2], and Mohd Sharifuddin Ahmad[3]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia
{mk,shamimi}@tmsk.uitm.edu.my
[2] Trauma and Emergency Department, Universiti Malaya Medical Center, Malaysia
rashidi@ummc.edu.my
[3] College of Information Technology, Universiti Tenaga Nasional, Malaysia
sharif@uniten.edu.my

Abstract. Triage is a decision-making process that classifies incoming patients for presentational urgency in Emergency Departments (EDs). There are issues with triage reliability in EDs, which we can be resolved through uniform application of a robust triage scale. However, the complex robust triaging knowledge is not easy to understand or recalled for timely decision-making. Therefore, we suggest the development of a knowledge-based triage decision support system to help triage officers to make correct and consistent triage decisions. Consequently, we pursued knowledge engineering to construct the models of the knowledge in order to make explicit the conceptualisation of the assumptions and constraints in triage decision-making. We regard task as a rationale basis for modelling the purposive domain knowledge. Consequently, the paper discusses the modelling of the domain knowledge to support the triage decision-making task. The triage decision-making task model is presented in a complementary paper. Together, the knowledge models can be viewed as meta models that provide the conceptual guiding principles for the consequent design of the triage decision support system.

Keywords: Knowledge Engineering, Domain Knowledge Modelling, Triage Assessment.

1 Introduction

Triage is the frontline assessment of incoming patients in Emergency Department (ED) where the triage officers are responsible for classification of presentational urgency. There are three issues with triage decision-making.

- The decision-making is not easy because it involves application of complex triaging knowledge to the discriminated and interpreted information elicited from patients, personal observations and physical examination [1].
- Inconsistency in triage decision-making is a major problem in ED [2]. Triage assessment under conditions of uncertainty by triage officers with different levels of expertise and experience, leads to different outcomes. Since triage classification

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_46, © Springer International Publishing Switzerland 2014

is correlated with patients' waiting times, it affects patents' satisfaction, and an under-triage can even place a patient at risk [3].

- The absence of a standard triage scale (c.f. in Malaysia) is another issue that affects triage reliability. It causes variable waiting times of patients and the assignment of a triage level can be potentially harmful if the scale is unreliable.

We think that the uniform application of a robust triage scale in EDs can resolve some of the above issues. A robust triage scale is capable of providing greater discrimination, better reliability and improved sensitivity and specificity [4, 5]. However, the robust triage decision-making is also complex. Complex knowledge is not easy to understand or recalled during rapid decision-making. Therefore, we suggest the development of a knowledge-based Triage Decision Support System (TDSS) to help the triage officers to make correct, consistent and timely triage decisions.

Consequently, we pursued knowledge engineering as a way to systematise the triage decision-making knowledge. The knowledge is modelled to make explicit the conceptualisation of the assumptions and constraints in triage decision-making.

We regard task as a rationale basis for modelling the purposive domain knowledge [6, 7]. The idea is to first model the triage decision-making task, then to focus on the modelling of the triage domain knowledge that is required to support the task. Together, the task and domain knowledge models can be viewed as meta models that provide the conceptual guiding principles for the consequent design of the triage decision support system.

This paper is structured as follows. Section 2 briefly describes the triage decision-making task knowledge model. In section 3, we give an example of a problem case presented in ED and the ensuing triage analysis. The example provides the background information to understand the discussion in the following sections. Sections 4 and 5 discuss the modelling of the triage domain knowledge and the model evaluation, respectively. Finally, we conclude in section 6.

2 Triage Decision-Making Task Model

Fig. 1 shows a simplified version of the triage task knowledge model presented in [8]. The model is adapted from the CommonKADS's Assessment generic task template [6], and captures the functional knowledge in triage decision-making. The inference structure in the original assessment task has been modified and reorganised to reflect the inferencing of knowledge in triage decision-making. The modifications include the introduction of a new inference: Sort. The Sort inference is used to prioritise the modifiers according to particular cases presented in ED. The prioritisation is based on the principles of emergency care [9].

The inferences are depicted by the ovals in the figure. The rectangular boxes connected to the inferences describe the knowledge and information used and produced by the inferences. For example, the *Sort-modifier* inference accepts *Specified-modifiers* as input and produces *Sorted-modifier* as output. In Table 1, we describe the key domain terms that feature in the triage task knowledge model.

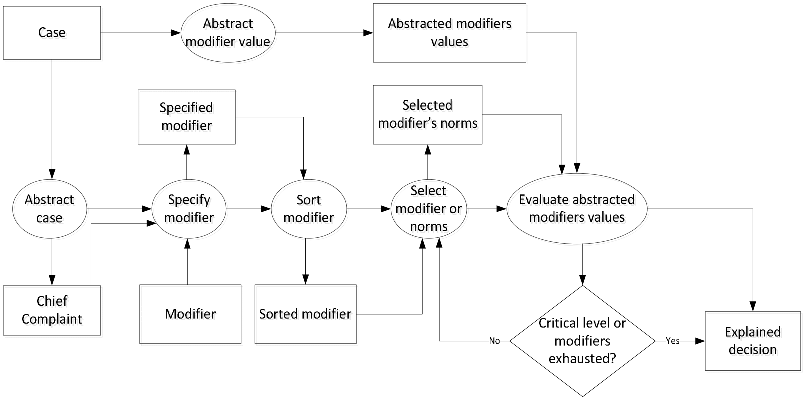


Fig. 1. A Simplified Triage Task Knowledge Model

Table 1. Key domain terms that support the triage decision-making task

Term	Description
Case	Knowledge or information gathered from the patient, which includes patient's oral history and clinical judgment. The clinical judgments involve objective and subjective measurements. The objective measurements comprise the vital sign (VS) readings, while the subjective measurements are based on observed signs and symptoms.
Chief Complaint	The most significant illness inferred from Case, which is based on Habboushe's guide [10].
Modifiers	Determinants used to determine the triage level of patients. For example, eleven modifiers feature in the Canadian Triage Acuity Scale (CTAS) [11]: Glasgow Coma Score (GCS), Respiratory distress, Hemodynamic stability, Dedicated presenting complaint, Mental health, Bleeding severity, Hypertension, Temperature, Pain, Mechanism of injury and Blood glucose.
Modifier value	The abstracted value assigned to a modifier. For example, GCS score is 13, and Pain is severe.

Modifier norms	A conditional rule of a modifier. Each modifier has more than one rule. An example of a simple GCS norm is
	IF GCS score is 13
	THEN triage level is II

Triage level	The triage level is an outcome of norm evaluation, which indicates the severity of a patient's clinical condition. The levels depend on the selected triage scale. For example, CTAS has five levels (I to V).
Critical level	The critical level refers to the most severe clinical condition. For example, level I in CTAS, i.e., when the patient necessitates immediate treatment.
Explained decision	The triage decision is justified by providing explanations that refer to the modifier norms that are applied to determine the triage level.

The flow line out of the condition box (diamond) at the bottom in Fig. 1 indicates the iterative consideration over the norms of the selected modifiers. The iteration continues until all modifier norms have been processed or a critical triage level is encountered. More details about the task model can be found in a complementary paper [8].

The task knowledge model serves as a rationale basis for modelling the triage domain knowledge, which is elaborated in section 4. But, first we give an example of a problem case presented in ED and the ensuing triage analysis in section 3. The example provides the background information that is essential to understanding of the modelling of the domain knowledge model in section 4.

3 A Problem Case

Case is a compulsory documentation that records information about a patient's background, illness and health problem including the kinds of treatment being received. It is the starting point of triage analysis. The case analysis leads to clinical judgment analysis, and is followed by the determination of the triage level.

We give an example of a simple case presented in ED for purpose of illustration and future reference. Tables 2, 3 and 4 summarise the analysis of the case, the analysis of the clinical judgment and the final analysis that determines the triage level, respectively.

Table 2. Case analysis

Element	**Description**
Patient	A 46 years old man … .
Oral history	The patient has had a headache for the last week, which is becoming worse. It is a constant ache, and is present over his entire head but is worst in his occipital area. Today, while the patient was at work at a dry cleaning establishment, the patient has an episode of confusion and syncope; so, he was brought to the hospital by his employer.
Past medical history	The patient has no known medical history, and denies medical problems. It is unknown whether the patient takes any medications. The patient has no known drug allergies.
Social history	The patient has been working at the dry cleaning establishment for the last eight months. He is not known to be a smoker or drinker. The patient denies drug use.
Chief complaint	Headache

The clinical judgement analysis is based on the chief complaint (the most significant illness abstracted from the case), information from physical examination (vital sign measurements), as well as information about any signs that are seen, heard, felt by the triage officer or any symptoms that are felt or experienced by the patient.

In our case, based on the chief complaint: Headache, the following modifiers were selected and ordered for assessment: Pain, GCS, Hypertension and Dedicated presenting complaint. The selection and ordering of the modifiers are based on the *Specify* and *Sort* inferencing knowledge. Table 3 captures the analysis of the clinical judgment that covers the assessment needs of the selected modifiers.

Table 3. Analysis of clinical judgment

Vital signs (VS)	
Temperature is 37.4°C, Pulse rate is 124 bpm, Respiratory rate is 20 breaths per minute, Blood pressure is 95/65, Air-oxygen saturation is 97%, Pain score 8 (locality is peripheral, acute) and Glucose level is 95.	
Other signs and symptoms	
General	The patient is confused. He is able to state his name in incomprehensive speech and recognises his employer, but he seems sleepy and nods off when not stimulated.
Head and neck	The patient's mucous membranes are moist. His neck is supple, although he seems uncomfortable when his neck is ranged. His sclerae are non-icteric.
Nervous system	The patient follows some commands and does not appear to have a focal neurologic deficit. His pupils are equal, round, and reactive to light. The patient does not complain when examined.
Cardiovascular	The patient's heart is regularly tachycardia, without murmurs, rubs or gallops.
Lung	The patient has no respiratory distress. His lungs are clear and equal. *Abdomen.* The patient is thin. His abdomen is soft and non-tender, with no organomegaly.
Extremities and skin	The patient's distal extremities are cool with poor capillary refill. He is not diaphoretic. He has no rashes.

Table 4 lists the key clinical measurements that were considered during the reasoning involving each of the selected modifier's norms, and the resulting triage level. Based on this analysis, the patient is triaged at level III, i.e., the lowest level inferred by the Pain and GCS modifier norms.

Table 4. Determination of triage level

Modifier	**Key considerations by norms**	**Triage level det. by modifier**
Pain	Pain score is 8, Locality is peripheral, acute	III
GCS	10 < GCS < 13 *(calculated value)*	III
Hypertension	90/60 < BP < 120/80 *(measured value)*	IV
Dedicated presenting complaint	No chest pain, No tearing, No injury, Element is upper body	IV

4 Modelling the Triage Domain Knowledge

The modelling of the triage domain knowledge is directed by its purposive mechanism. In this regards, the triage domain knowledge model characterises the domain resources required to support the triage decision-making task. Accordingly, we followed Annamalai and Sterling's method for constructing purposive ontologies [7], in order to conceptualise the triage domain knowledge. However, we made modifications to the steps to reflect the task based approach we have taken to model the domain knowledge. We describe the steps below.

a) Establish the purpose and use of the domain knowledge model. In our case, the model purports to conceptually guide the consequent design of the triage decision support system. In view of that, the triage decision-making task knowledge model (Fig. 1) is constructed in advance to serve as the frame of reference for the conceptualisation of the triage domain knowledge.
b) Identify the preliminary concepts that stand for the domain knowledge resources supporting the triage decision-making task. We begin by recognising the distinct concepts required for describing the result of triage analysis (see Tables 2 – 4). The preliminary concepts are Case, Vital sign, Observed sign and symptom, Chief complaint, Modifier, Triage level, and so on. *Note:* The preliminary concepts serve as links for structuring additional concepts into the model.
c) Sketch an outline of the preliminary concept model.
d) Identify additional concepts required to detail and refine the model. New concepts affiliated with the preliminary concepts are organised and related with other concepts in the model using bottom-up and middle-out processes. Examples of additional concepts included later in the model are: Norms, Organ system, Concentration of emergency, Units of measure, and so on.
e) Structure and relate the identified concepts into the domain knowledge model (Fig. 2). For example, the seed concepts in the model are Case and Clinical judgment (see Tables 2 and 3). The concepts that are immediately related to these seed concepts through aggregation and composition relationships are shaded in the model.
f) Evaluate and make the necessary changes to the domain knowledge model until satisfied (repeat steps (d) – (f)). The model evaluation is discussed in section 5.

Fig. 2 describes a partial hierarchy of the concepts in the triage domain knowledge model. Individuals in this conceptualisation are concept terms, which are constrained by properties and relations. The model is developed iteratively until we are satisfied, i.e., when the domain knowledge resources supporting the decision-making task can be characterised using the terms in the model. The model is elaborated from triage related books, articles, glossaries and by consulting the domain experts. The model also unifies existing taxonomies with partial characterisation of domain through analysis and synthesis.

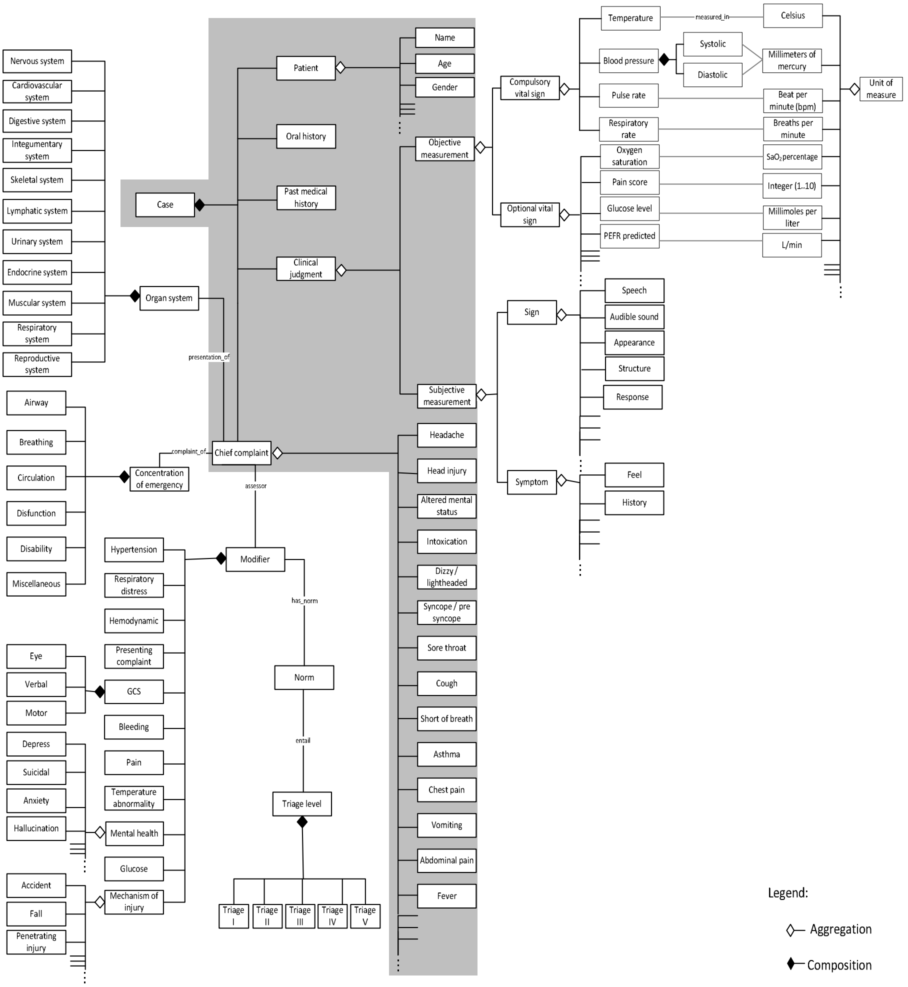


Fig. 2. Triage Domain Knowledge Model

5 Model Evaluation

The formative evaluation of the knowledge model is necessary to ensure the quality of the model being developed. The evaluation will be based on a prescribed set of evaluation criteria, namely Consistency, Completeness, Conciseness, Competency, Extensibility and Expressiveness. These criteria espouse the general design principles of domain knowledge models [12, 13].

- *Consistency* criterion evaluates the logical and structural element of the model. The analysis can be divided into two parts: Conceptual integrity and Collective consistency [12].

In conceptual integrity, we checked if the individual concepts in the model correspond to specific entities in the area of the domain knowledge, i.e., we verify whether the concepts and related properties are sensible vocabulary. The evaluation is made with respect to the information used to articulate the concept's structure.

In collective consistency, we examined the coherence between the concepts verifying if the logical relationships that bind these concepts in the model reflect the dependencies between their corresponding entities in the area of knowledge.

- *Completeness* criterion checks whether the model has covered all the required concepts. In practice, completeness is hard to achieve since models are by nature, incomplete. Therefore, we strived for functional completeness [12, 14] and/ or epistemological completeness [15].

  Under functional completeness, we assessed the functional adequacy of the domain knowledge model in the context of its purpose of design. In our case, the functional needs are exemplified in the triage decision-making task knowledge model. Therefore, we checked to ensure that the presented concepts are able to cover the characterisation of the domain knowledge to support the triage decision-making.

  Under epistemological completeness, we check the concepts in the model for exhaustiveness/ incomplete classification, granularity or level of detailness involving the properties and relations of the concepts.

- *Competency* criterion checks to ensure that a model is capable of supporting the purpose of its design and use (c.f. Completeness) [12]. In principle, both functional and epistemological completeness are the necessary basis for evaluating the 'competence' of a model. We checked whether the concepts in the model can definitely convey the relevant contents of the domain knowledge resources, so that they can be 'competently' utilised to support the triage decision-making task. In a sense, the competency evaluation assesses the potential use of the model.

- *Conciseness* criterion is used to evaluate the relevancy or relatedness of presented concept with respect to their needs or requirements [14]. In practice, the analysis actually does the opposite, i.e., remove redundant and irrelevant concepts that are present in the model. Unnecessary and unwanted concepts do not add value to the model and sometimes can lead to inconsistencies in the model. The point is to checks whether the concepts and related properties and relations are relevant for modelling the triage domain knowledge, and remove concepts, properties and relations whose presence in the model cannot be justified [12]. We carried out this analysis with the help domain experts who expressed their agreement or disagreement to the presented concepts.

- *Extensibility* criterion checks to ensure that the concepts are structured in the manner that facilitates future extension of the model. A model must be easily extendable to allow for incremental additions, modifications and deletions without having to reorganise the existing structure, i.e., degenerating its present state of being. In our case, the conceptual description is objectified, which facilitates ease of expansion. Moreover, the hierarchical organisation of the concepts using the aggregation and composition relations provides a framework for ordered representation of the generalised concepts that is easy to change, which also scale easily.

- *Expressiveness* is a consequential criterion of a model. Since the model must be able effectively express the intended meaning of the concepts, the formalism in which a knowledge model is represented is a tangible factor in determining its expressivity. In our case, the domain knowledge model is a simplified conceptual structure represented using the Unified Modelling Language (UML), which is familiar to most software developers (an intended user). *Note*: The purpose of the model is to conceptually guide the consequent design of the triage decision support system.

  Much of the concepts in the model are related using the standard aggregation and composition relationships. There are only few user-defined associative relations in the model. Therefore, the UML representation is easy to understand without requiring additional information (and so, can be applied consistently). This facilitates the evaluation of the model, and its eventual use in design and development.

6 Conclusion

This paper discusses a task-based approach for the modelling of the triage domain knowledge. The purposive domain knowledge model is conceptualised in the context of the triage decision-making task. The triage domain knowledge model was conceived to serve as the rationale basis of the design and development of a triage decision-support system. The construction of domain knowledge model is adapted from a purposive ontology development method as described in section 3. The iterative development of the domain knowledge model includes its formative evaluation based on general design principles embodied in a set of criteria. The resulting model fulfils the condition of an elementary model, which has purpose of design, supports the ED triage community of practice, and proposes a method for utilisation [16]. We hope to extend the model into a mature model as part of future work.

Acknowledgements. This work was supported by the Malaysian Ministry of Education's Exploratory Research Grant Scheme (reference number: ERGS/1/2013/ ICT07/UITM/02/02).

References

1. Larkin, G.L., Marco, C.A., Abbott, J.T.: Emergency Determination of Decision-making Capacity: Balancing Autonomy and Beneficence. The Emergency Department, Academic Emergency Medicine 8, 282–284 (2001)
2. Wuerz, R.C., Fernandes, C.M.B., Alarcon, J.: Inconsistency of Emergency Department Triage. Annals of Emergency Medicine 32, 431–435 (1998)
3. Halim, S., Annamalai, M., Ahmad, M.S., Ahmad, R.: A Conceptualisation of an Agent-Oriented Triage Decision Support System. In: Lukose, D., Ahmad, A.R., Suliman, A. (eds.) KTW 2011. CCIS, vol. 295, pp. 272–282. Springer, Heidelberg (2012)

4. Travers, D.A., Waller, A.E., Bowling, J.M., Flowers, D., Tintinalli, J.: Five-level Triage System More Effective than Three-level in Tertiary Emergency Department. Journal of Emergency Nursing 28, 395–400 (2002)
5. Masturzo, P., Regolo, R., Ferro, G., Nardi, G., Orazi, D., Maggi, V.: Sensitivity and Specificity of a Triage Score Dedicated to Trauma Patients in a Tertiary-level Hospital: Preliminary Results. Critical Care 11(2), 354 (2007)
6. Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., de Velve, W.V., Wielinga, B.: Knowledg Engineering and Management: The CommonKADS Methodology. The MIT Press, Cambridge (2000)
7. Annamalai, M., Sterling, L.: Guidelines for Constructing Reusable Domain Ontologies. In: The AAMAS 2003 Workshop on Ontologies in Agent Systems, Melbourne, Australia (2003)
8. Halim, S., Annamalai, M., Ahmad, R., Ahmad, M.S.: Task Knowledge Model for Triage Decision-Support. In: 5th International Conference on Knowledge Engineering and Ontology Development, Vilamoura, Portugal (to appear, 2013)
9. Digna, R.K., Johan, G.B.: Advanced Trauma Life Support: ABCDE from a Radiology Point of View. Emergency Radiology 14(3), 135–141 (2004)
10. Habboushe, J.: The Basics of Emergency Medicine: A Chief Complaint Based Guide. EMRA Publications (2012)
11. Murray, M., Bullard, M., Grafstein, E.: Revisions to the Canadian Emergency Department Triage and Acuity Scale Implementation Guidelines. Canadian Journal of Emergency Medicine (2004)
12. Annamalai, M.: Formative Evaluation of Ontologies for Information Agents. In: The Conference on Computer Science, Technology and Networking, Shah Alam, Malaysia (2005)
13. Gomez-Perez, A.: Evaluation of Ontologies. International Journal of Intelligent Systems 16(3), 391–401 (2001)
14. Mehmood, K., Cherfi, S.S.-S.: Evaluating the Functionality of Conceptual Models. In: Heuser, C.A., Pernul, G. (eds.) ER 2009. LNCS, vol. 5833, pp. 222–231. Springer, Heidelberg (2009)
15. Ribbert, M., Niehaves, B., Dreiling, A., Holten, R.: An Epistemological Foundation of Conceptual Modeling. In: The 12th European Conference on Information Systems, Turku, Findland (2004)
16. Thalheim, B.: The Conception of the Model. In: Abramowicz, W. (ed.) BIS 2013. LNBIP, vol. 157, pp. 113–124. Springer, Heidelberg (2013)

A Survey of High Level Synthesis Languages, Tools, and Compilers for Reconfigurable High Performance Computing

Luka Daoud[1], Dawid Zydek[1], and Henry Selvaraj[2]

[1] Department of Electrical Engineering, Idaho State University, Pocatello, ID, USA
{daouluka,zydedawi}@isu.edu

[2] Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, NV, USA
henry.selvaraj@unlv.edu

Abstract. High Level Languages (HLLs) make programming easier and more efficient; therefore, powerful applications can be written, modified, and debugged easily. Nowadays, applications can be divided into parallel tasks and run on different processing elements, such as CPUs, GPUs, or FPGAs; for achieving higher performance. However, in the case of FPGAs, generating hardware modules automatically from high level representation is one of the major research activities in the last few years. Current research focuses on designing programming platforms that allow parallel applications to be run on different platforms, including FPGA. In this paper, a survey of HLLs, tools, and compilers used for translating high level representation to hardware description language is presented. Technical analysis of such tools and compilers is discussed as well.

Keywords: High Level Synthesis, Compilers, FPGA, C-to-VHDL.

1 Introduction

Today, the trend in High Performance Computing (HPC) is to run applications in parallel on different processing elements. Applications can be divided into multiple tasks and executed simultaneously. Performance of such parallel processing systems has increased significantly over the past few years and that trend continues till date. The improvement is possible by implementing the multi-core versions of conventional CPUs (Central Processing Units), and by hybrid computing platforms with accelerators, such as FPGAs (Field Programmable Gate Arrays) or GPUs (Graphics Processing Units) [1,2].

In general, applications and simulations [3] are not well suited for executing exclusively on accelerators. Some portions of applications have extensive parallelism, suitable for FPGAs or GPUs; others are inherently serial or have extensive control flow that make them better suited for a CPU or again for an FPGA. Such device-oriented application partitioning increases the overall system performance.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_47, © Springer International Publishing Switzerland 2014

Hybrid HPC is a promising approach to increase the performance of supercomputers. It combines computing platforms, such as CPUs, GPUs, and FPGAs; in order to attain higher performance. FPGAs offer energy efficiency and higher throughput for portions of applications characterized by simple data objects and extensive parallelism. GPUs outperform FPGAs for streaming and floating-point applications; and for applications requiring high memory bandwidth. CPUs are optimized for serial processing. Combining these computing platforms ensures HPC and decreases energy consumption. The main challenge in recent HPC software is to design a solution that allows using CPU (as a main processing unit), GPUs, FPGAs, and other accelerators. Moreover, the software should be able to decide which portion of the application is suitable for which computing platform, considering higher performance and energy efficiency (Fig. 1).

In order to implement application code on an FPGA, the code should be written in Hardware Description Language (HDL), like e.g. in [4] or [5]. Due to rapid increase of complexity in the systems, researchers and engineers moved from Register Transfer Level (RTL) design to high level design, seeking better productivity in less time and with lower cost. Therefore, this paper provides a survey on current and recent High Level Synthesis (HLS) tools, languages, and compilers for reconfigurable systems. These HLS compilers are categorized in this paper based on the input language. This paper is a first step in developing a new tool that translates High Level Language (HLL) to a code recognizable by variety of computing hardware, such as CPUs, GPUs, FPGAs, DSPs (Digital Signal Processors), or others. The paper is organized as follows: The next section compares FPGAs to CPUs and GPUs. A survey on HLS, languages, and compilers is presented in Section 3; and finally Section 4 concludes the paper.

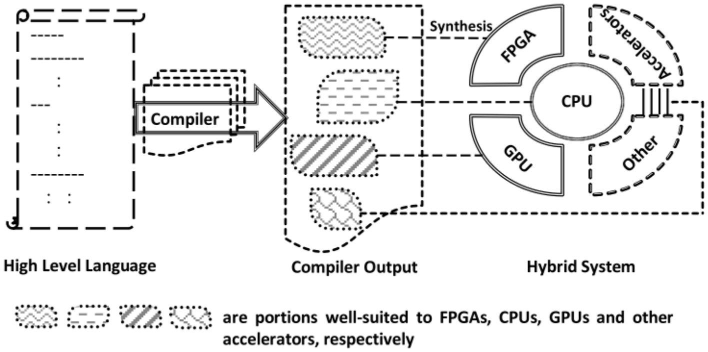


Fig. 1. A compiler compiling an application-code for CPUs, GPUs, FPGAs, and other accelerators

2 FPGAs vs. CPUs and GPUs

FPGAs and/or GPUs are used in hybrid computing systems to accelerate the computing processes [1]. Hence, HPC can be obtained. In HPC systems, the FPGA acts as a configurable co-processor to a CPU, where FPGA executes intensive computational parts of the code. Similarly, GPU processes large blocks of data in parallel to increase performance. Although FPGAs run at frequencies expressed in MHz while CPUs run at few GHz, very often FPGAs outperform CPUs. The reasons behind that are:

- For each specific task a dedicated circuit is implemented in the FPGA,
- Designers exploit the parallelism and pipeline of the circuit implemented on the FPGA,
- FPGAs offer huge memory bandwidth through configurable logic.

Besides higher performance offered by FPGAs, they provide lower power consumption in comparison to CPUs. In HPC systems, GPUs are intensively used for numerous scientific applications to achieve higher performance by off-loading the most intensive computing portions of the application to the GPUs. GPUs often outperform FPGAs for streaming applications. They usually have a higher floating-point performance and memory bandwidth than FPGAs. However, according to [6], FPGAs present better computing capability for applications characterized by:

- Relatively simple data objects,
- Relatively simple arithmetic operations,
- Smooth implementation using pipelined processing structures,
- Extensive data-parallelism,
- Regular and simple control structures.

3 HLS Languages, Tools, and Compilers

In this section, we analyze HLS tools and programming languages for FPGAs. The tools enhance the portability and scalability of applications. Moreover, they optimize performance and design efforts as well. Most of these tools are based on C/C++ programming language, because the language is well known for both software and hardware engineers. Fig. 2. shows the flow of generating RTL from HLL.

3.1 HLS from C/C++ Programming Language

Hardware-C [7] is one of the first HLS languages that uses the C programming language. It supports parallel processes that communicate together through either port passing or message passing techniques. However, it cannot represent arbitrary serial-parallel structures and it has different syntax from the original C for several constructs. Handel-C [8] is one of the HLLs based on C language,

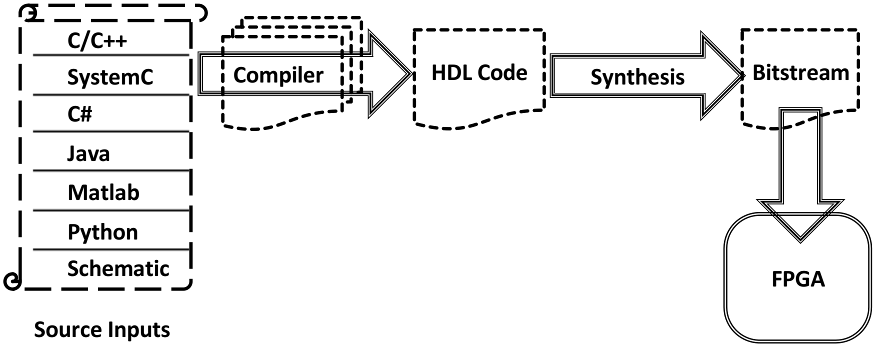


Fig. 2. The flow of generating RTL from HLL

where the commands are written one by one and they are executed sequentially. Handel-C targets low level hardware where the commands can be executed in FPGA. In order to get benefit of parallel execution, some keywords are used in the code. Therefore, performance benefits can be attained by using parallelism. The overall program structure of Handel-C is little different from conventional C. The program structure consists of one or more main functions, each associated with a clock. Thus parts of the program can be run at different speeds. When a code is written in Handel-C, the programmer should be aware of the hardware implementations in order to get the benefit of parallelism. Similar to Handel-C, Hyden-C [9] is a framework of optional annotations to enable designers to describe design-constraints and to direct source-level transformations such as scheduling and resource allocation. The main difference between Handel-C and Hyden-C is that Hyden-C, like VHDL, is component-based. Therefore, designers can describe their designs as a set of distinct components that are developed independently and then connect them together.

Many HLS tools are designed for application domains. For example, Trident [10] is a compiler that accepts C code extracting the parallelism and the possibility of pipeline implementations from the code; and generates the corresponding circuits in reconfigurable logic. Trident compiler is mainly designed for floating-point applications. GAUT [11] is an academic HLS tool dedicated to DSP applications. The GAUT tool converts a C function into a pipelined architecture consisting of a processing unit, memory unit, communication, and multiplexing unit. Also, Streams-C [12] and Impulse-C (derived from Streams-C) [13] are compilers that support stream-oriented computation on FPGA-based parallel processors, where data parallelism can be effectively mapped onto the FPGA.

In addition, many of the free and open source online compilers can be used to convert the C code (HLL) to HDL. C to Verilog [14] is an online compiler that translates the C function into a hardware-module interface. Although this compiler uses most of the C language features, there are some limitations in the C code that are not acceptable by the tool, e.g. recursive functions, structures, pointers to functions, and library function calls (printf, malloc, etc). These structures (limitations) cannot be represented in hardware. FpgaC [15] is a compiler for a subset of the C programming language. It produces digital circuits that execute the compiled programs.

Since hybrid systems provide higher performance, some compilers target hardware and software systems by translating specific parts of a larger C program into hardware; and the rest of the program is executed on a traditional CPU. NAPA C [16] is a hardware-software and co-synthesis compiler that generates a C program targeting hybrid RISC CPUs and FPGAs. Similarly, Nimble [17] is a framework that automatically compiles system-level applications specified in C to the code executable on the combined CPU and FPGA architectures. Also, CHiMPS [18] is a C-based compiler for hybrid CPU-FPGA computing platform. Similar to CHiMPS, CASH [19] is a compiler that targets the hybrid System on Chips (SoCs). LegUp [20] is an open source HLS tool that automatically compiles a C program to target a hybrid FPGA-based software/hardware system. The program can be divided into program segments, some program segments are executed on an FPGA-based MIPS CPU and other program segments are automatically synthesized into FPGA circuits. These circuits communicate and work together with the CPU. ROCCC [21] is an open source compiler that accepts a strict subset of C and generates VHDL. In order to be used efficiently, the entire software program is not translated into hardware – ROCCC focuses on the critical regions of software, e.g. regions containing loops performing extensive computation on large amounts of data. The Nios II C-to-Hardware Acceleration (C2H) compiler [22] is a tool that allows the designer to create custom hardware accelerators directly from C code. Altera's C2H allows partitioning C functions into hardware sets that can be executed on a Nios II CPU. By using the C2H compiler, an algorithm in C targeting a Nios II CPU can be quickly converted to a hardware accelerator implemented on an Altera's FPGA.

There are also C++ language compilers that can be used to generate HDL. A Stream Compiler (ASC) [23] is a C++ library allowing the designers to optimize the hardware implementation on the algorithm level, architecture level, arithmetic level, and gate level; all within the same C++ program. ASC code is compiled to produce hardware netlist circuit. Catapult C [24] is a subset of C++ with no extensions. It takes ANSI C/C++ or SystemC [25] as input and generates RTL code targeting FPGAs or ASICs. The code that can be compiled from Catapult C may be very general and may result in many different hardware implementations. AutoPilot [26] is one of the most recent HLS tools. It automatically generates efficient RTL code from high level representations. Autopilot accepts three kinds of standard C-based design entries: C, C++, and SystemC. AutoPilot is an advanced compiler capable of carrying out efficient power optimization using clock gating and power gating. It also supports pipeline to improve the system performance. Hence, it can target a wide range of applications.

In addition, DEFACTO [27] and Carte [28] are compilers that accept C and Fortran as input languages. DIME-C [29] is a C based compiler that translates a DIME-C code into VHDL. However, like C to Verilog [14], not all elements in C languages are supported in DIME-C, e.g. pointers, structures, switch statements, etc. Other C-based compilers include Bash-C [30], Mitrion-C [31], SpecC [32], SPC [33], or SPARK [34]. Additional commercial tools and early compilers can be found in [35].

3.2 HLS from Non-C/C++ Programming Language

MATlab Compiler for Heterogeneous computing systems (MATCH) [36] allows users to develop efficient codes for distributed, heterogeneous, and reconfigurable computing systems. MATCH takes MATlab programs and automatically maps them onto a hybrid computing environment consisting of embedded CPUs, DSPs, or FPGAs.

MyHDL [37] is an open source Python package that allows the designer using Python to generate HDL. The Python code is converted to Verilog and VHDL.

JHDL (Just-Another Hardware Description Language)[38] is a design tool for reconfigurable systems that focuses mainly on designing circuits through an object oriented approach. The main use of JHDL is to create digital circuits for implementation using FPGAs. JHDL can be used with any standard Java 1.1 distribution without language extensions. Also, based on Java source input, Sea Cucumber (SC) [39] is a synthesizing compiler for FPGAs that accepts Java class files as input and then generates circuits. Users write circuit descriptions exposing coarse-level parallelism as concurrent threads. SC then analyzes the body of each thread and uses compiler and circuit optimization techniques to extract fine-grained parallelism. Afterwards, SC compiler is executed to generate an Electronic Design Interchange Format (EDIF) netlist; and the Xilinx place and route software is called to create a bitstream from the synthesized EDIF netlist.

Kiwi [40] is a compiler based on C#. The Kiwi compiler accepts common intermediate language output from either the .NET or Mono C# compilers and generates Verilog RTL.

Pebble [41] is a language for parameterized and reconfigurable hardware design. Pebble has a simple block-structured syntax. Designers can easily define the number of pipeline stages using the parameters in a Pebble program. The main objective of Pebble is to support the development of designs involving run-time reconfiguration.

There are also other HLS compilers that use different languages. For example, Esterel [42] is a synchronous programming language designed to program reactive systems (systems that react continuously to their environment). Another example is BlueSpec compiler [43] that works based on Bluespec System Verilog – a language used in the design of electronic systems.

3.3 Schematics-Based HLS

Schematics such as LabVIEW and MATlab are also used to program FPGAs. LabVIEW is one of the schematic tools that targets FPGAs. The National Instruments (NI) LabVIEW FPGA module extends the LabVIEW graphical development platform to target FPGAs on NI reconfigurable I/O hardware. Since LabVIEW represents parallelism and data flow, it is suitable for FPGA programming [44]. In addition, designers also can use MATlab to design and simulate their algorithms by Simulink and Stateflow, then MATlab generates VHDL or Verilog code for FPGAs using HDL Coder [45]. Another tool is Altium Designer [46]. It is an electronic design automation software package for printed circuit board, FPGA, and embedded software design.

3.4 HLS Based on Programming Models for GPUs

Parallel computing platforms and programming models for GPUs are subject of very intense research. CUDA (Compute Unified Device Architecture) and OpenCL (Open Computing Language) are parallel programming models that address the higher interest in GPUs; moreover, they have recently expanded their capabilities beyond GPUs.

CUDA is a parallel computing platform and programming model created by NVIDIA and implemented on their GPUs. FCUDA [47] is a framework to convert CUDA code to RTL suitable for FPGAs. The transformation process from CUDA to RTL is done in two phases. First, FCUDA transforms the single-program-multiple-data CUDA code into C code for AutoPilot [25] with annotated coarse-grained parallelism. Then, the AutoPilot maps the marked parallelism onto parallel cores and generates the corresponding RTL description. Afterwards, synthesis and programming of FPGA is done. The main goal of the FCUDA is to convert thread blocks into C functions. FCUDA combines the CUDA programming model with a HLS tool (AutoPilot) to efficiently implement CUDA code on FPGA.

Another parallel programming framework for writing programs that are executed across heterogeneous platforms is OpenCL [48]. SOpenCL [49] is an OpenCL-based FPGA synthesis tool. It generates hardware circuits and SoC systems from OpenCL programs. The output of SOpenCL is a pure C function which is converted to a hardware circuit in a form of synthesizable HDL.

On the other hand, Altera enables the designers to run OpenCL code on Altera's FPGAs [50]. Compiling an OpenCL code to FPGA by Altera's solution is the process of converting an OpenCL C code into FPGA bitstream that allows programming the FPGA. The compilation process has two phases: the OpenCL code is compiled into intermediate hardware format code, and then compiled into an FPGA bitstream. Therefore, each OpenCL code is converted into custom hardware representing the data flow circuit. Mapping multithreaded functions to FPGA can be done simply by replicating hardware (inefficient – waste of resources) or by using pipeline parallelism (more efficient mapping).

4 Final Remarks

The trend in HPC is to increase the number of processing elements using heterogeneous computing platforms, in order to provide higher performance and increase energy efficiency. Since the complexity in applications and systems has been increasing, designers move to high level representation to improve the product quality in less time and with lower cost.

This paper is a survey of HLS tools and compilers that accept HLL code and generate HDL or bit-stream files for FPGAs. These HLS tools and compilers have been presented according to the input source code. Most current and recent compilers have been presented. Also, compilers that convert a code for GPUs, written in CUDA or OpenCL, to RTL form have been demonstrated in this paper. This work has been presented as the first step to design a compiler capable of compiling and analyzing code for heterogeneous systems combining CPUs, GPUs, and FPGAs. As a future work, we plan to design such a compiler that will use intelligent techniques to select the best computing platform for all portions of the code, in order to increase performance. Our future work will also focus on improving efficiency of FPGAs for floating-point calculations.

References

1. Liu, B., Zydek, D., Selvaraj, H., Gewali, L.: Accelerating High Performance Computing Applications Using CPUs, GPUs, Hybrid CPU/GPU, and FPGAs. In: Proceedings of the 13th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2012), pp. 337–342 (2012), doi:10.1109/PDCAT.2012.34
2. Zydek, D., Selvaraj, H., Borowik, G., Luba, T.: Energy Characteristic of Processor Allocator and Network-on-Chip. International Journal of Applied Mathematics and Computer Science 21(2), 385–399 (2011), doi:10.2478/v10006-011-0029-7
3. Chmaj, G., Zydek, D.: Software development approach for discrete simulators. In: Proceedings of the 21st International Conference on Systems Engineering (ICSEng 2011), pp. 273–278. IEEE Computer Society Press (2011), doi:10.1109/ICSEng.2011.56
4. Zydek, D., Selvaraj, H., Gewali, L.: Synthesis of Processor Allocator for Torus-based Chip Multiprocessors. In: Proceedings of the 7th International Conference on Information Technology: New Generations (ITNG 2010), pp. 13–18. IEEE Computer Society Press (2010), doi:10.1109/ITNG.2010.145
5. Zydek, D., Selvaraj, H.: Hardware Implementation of Processor Allocation Schemes for Mesh-based Chip Multiprocessors. Microprocessors and Microsystems 34(1), 39–48 (2010), doi:10.1016/j.micpro.2009.11.003
6. Chase, J., Nelson, B., Bodily, J., Wei, Z., Lee, D.: Real-time Optical Flow Calculations on FPGA and GPU Architectures: A Comparison Study. In: Proceedings of the 16th International Symposium on Field-Programmable Custom Computing Machines, FCCM 2008, pp. 173–182 (2008)
7. Ku, D.C., De Micheli, G.: Hardware C - A Language for Hardware Design. Technical report, DTIC Document (1988)
8. Aubury, M., et al.: Handel-C Language Reference Guide. Computing Laboratory. Oxford University, UK (1996)
9. Coutinho, J., Luk, W.: Source-directed Transformations for Hardware Compilation. In: Proceedings of the International Conference on Field-Programmable Technology (FPT), pp. 278–285. IEEE (2003)

10. Tripp, J., et al.: Trident: An FPGA Compiler Framework for Floating-Point Algorithms. In: Proceedings of the International Conference on Field Programmable Logic and Applications, pp. 317–322 (2005)
11. GAUT- High-Level Synthesis Tool From C to RTL (May 2013), http://hls-labsticc.univ-ubs.fr
12. Gokhale, M., Stone, J., Arnold, J., Kalinowski, M.: Stream-Oriented FPGA Computing in the Streams-C High Level Language. In: 2000 IEEE Proceedings of the Symposium on Field-Programmable Custom Computing Machines, pp. 49–56 (2000)
13. http://www.impulseaccelerated.com/ReleaseFiles/Help/ImpulseCUserGuide.pdf: (May 2013)
14. C to Verilog (May 2013), http://www.c-to-verilog.com
15. FpgaC Compiler (May 2013), http://www.utb.edu/vpaa/csmt/cis/Pages/FPGAc.aspx
16. Gokhale, M., Stone, J.: NAPA C: Compiling for a Hybrid RISC/FPGA Architecture. In: Proceedings of the IEEE Symposim on FPGAs for Custom Computing Machines, pp. 126–135 (1998)
17. Li, Y., et al.: Hardware-Software Co-Design of Embedded Reconfigurable Architectures. In: Proceedings of the 37th Annual Design Automation Conference, pp. 507–512 (2000)
18. Putnam, A., et al.: CHIMPS: A C-Level Compilation Flow for Hybrid CPU-FPGA Architectures. In: 2008 FPL, Proceedings of the International Conference on Field Programmable Logic and Applications, pp. 173–178 (2008)
19. Budiu, M., Goldstein, S.C.: Compiling Application-Specific Hardware. In: Glesner, M., Zipf, P., Renovell, M. (eds.) FPL 2002. LNCS, vol. 2438, pp. 853–863. Springer, Heidelberg (2002)
20. Canis, A., et al.: LegUp: High-Level Synthesis for FPGA-based Processor/Accelerator Systems. In: Proceedings of the 19th ACM/SIGDA International Symposium on Field Programmable Gate Arrays, pp. 33–36 (2011)
21. Villarreal, J., Park, A., Najjar, W., Halstead, R.: Designing Modular Hardware Accelerators in C with ROCCC 2.0. In: Proceedings of the 18th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), pp. 127–134 (2010)
22. http://www.altera.com/literature/ug/ug_nios2_c2h_compiler.pdf (May 2013)
23. Mencer, O.: ASC: A Stream Compiler for Computing with FPGAs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 25(9), 1603–1617 (2006)
24. Mentor Graphics, Catapult C Synthesis (May 2013), http://www.mentor.com
25. Initiative, Open SystemC: SystemC 2.0. 1 Language Reference Manual. Revision 1(1177), 95118–3799 (2003)
26. Zhang, Z., Fan, Y., Jiang, W., Han, G., Yang, C.: AutoPilot: A platform-based ESL Synthesis System. In: High-Level Synthesis, pp. 99–112. Springer (2008)
27. Bondalapati, K., et al.: DEFACTO: A Design Environment for Adaptive Computing Technology. Springer (1999)
28. Poznanovic, D.S.: Application Development on the SRC Computers, Inc. Systems. In: Proceedings of the 19th IEEE International Symposium on Parallel and Distributed Processing, pp. 1–10 (2005)
29. Park, S., Shires, D., Henz, B.: Reconfigurable Computing: Experiences and Methodologies. Technical report, DTIC Document (2008)

30. Yamada, A., et al.: Hardware Synthesis with the Bach System. In: Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, ISCAS 1999, vol. 6, pp. 366–369 (1999)
31. Mitrionics AB: Mitrion Users Guide. Technical report, Mitrionics (2008)
32. Domer, R.: The SpecC System-Level Design Language and Methodology, Part 1, Parts 1 & 2. In: Embedded Systems Conference (2001)
33. Weinhardt, M., Luk, W.: Pipeline Vectorization. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 20(2), 234–248 (2001)
34. Gupta, S., Dutt, N., Gupta, R., Nicolau, A.: SPARK: A High-Level Synthesis Framework for Applying Parallelizing Compiler Transformations. In: Proceedings of the 16th International Conference on VLSI Design, pp. 461–466 (2003)
35. Cong, J., et al.: High-Level Synthesis for FPGAs: From Prototyping to Deployment. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 30(4), 473–491 (2011)
36. Banerjee, P., et al.: A MATLAB Compiler for Distributed, Heterogeneous, Reconfigurable Computing Systems. In: Proceedings of IEEE Symposiumon Field-Programmable Custom Computing Machines, pp. 39–48 (2000)
37. MyHDL - From Python to Silicon: myhdl.org (May 2013)
38. Bellows, P., Hutchings, B.: JHDL - An HDL for Reconfigurable Systems. In: Proceedings of the IEEE Symposium on FPGAs for Custom Computing Machines, pp. 175–184 (1998)
39. Tripp, J.L., Jackson, P.A., Hutchings, B.L.: Sea Cucumber: A Synthesizing Compiler for FPGAs. In: Glesner, M., Zipf, P., Renovell, M. (eds.) FPL 2002. LNCS, vol. 2438, pp. 875–885. Springer, Heidelberg (2002)
40. Greaves, D., Singh, S.: Using C# Attributes to Describe Hardware Artefacts within kiwi, Specification, Verification and Design Languages. In: Forum on Specification, Verification and Design Languages, FDL 2008, pp. 239–240 (2008)
41. Luk, W., McKeever, S.: Pebble: A Language For Parametrised and Reconfigurable Hardware Design. In: Hartenstein, R.W., Keevallik, A. (eds.) FPL 1998. LNCS, vol. 1482, pp. 9–18. Springer, Heidelberg (1998)
42. Berry, G., Gonthier, G.: The Esterel Synchronous Programming Language: Design, Semantics, Implementation. Science of Computer Programming 19(2), 87–152 (1992)
43. BlueSpec (May 2013), `http://www.bluespec.com`
44. National Instruments LabVIEW (May 2013), `http://www.ni.com/labview/fpga`
45. FPGA Design and Codesign (May 2013), `http://www.mathworks.com/fpga-design`
46. Altium Designer (May 2013), `http://en.wikipedia.org/wiki/Altium_Designer`
47. Papakonstantinou, A., et al.: FCUDA: Enabling Efficient Compilation of CUDA Kernels onto FPGAs. In: Proceedings of the 7th IEEE Symposium on Application Specific Processors, SASP 2009, pp. 35–42 (2009)
48. khronos Group (May 2013), `http://www.khronos.org`
49. Owaida, M., Bellas, N., Daloukas, K., Antonopoulos, C.: Synthesis of Platform Architectures from OpenCL Programs. In: Proceedings of the 19th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), pp. 186–193 (2011)
50. Implementing FPGA Design with the OpenCL Standard (May 2013), `http://www.altera.com/literature/wp/wp-01173-opencl.pdf`

Matrix Multiplication in Multiphysics Systems Using CUDA

Dawid Krol[1], Dawid Zydek[1], and Henry Selvaraj[2]

[1] Department of Electrical Engineering, Idaho State University, Pocatello, ID, USA
{kroldawi,zydedawi}@isu.edu

[2] Department of Electrical and Computer Engineering, University of Nevada,
Las Vegas, NV, USA
henry.selvaraj@unlv.edu

Abstract. Multiphysics systems are used to simulate various physics phenomena given by Partial Differential Equations (PDEs). The most popular method of solving PDEs is Finite Element method. The simulations require large amount of computational power, that is mostly caused by extensive processing of matrices. The high computational requirements have led recently to parallelization of algorithms and to utilization of Graphic Processing Units (GPUs). To take advantage of GPUs, one of GPU programming models has to be used. In this paper, CUDA model developed by nVidia is used to implement two parallel matrix multiplication algorithms. To evaluate the effectiveness of these algorithms, several experiments have been performed. Results have been compared with results obtained by classic Central Processing Unit (CPU) matrix multiplication algorithm. The comparison shows that matrix multiplication on GPU significantly outperforms classic CPU approach.

Keywords: CUDA, Matrix Multiplication, Multiphysics Simulations, libMesh.

1 Introduction

Multiphysics simulation is a very complex and time consuming process, which incorporates significant number of physical phenomena described by Partial Differential Equations (PDEs). The phenomena depict various areas of science, like e.g. solid mechanics, heat transfer, automobile systems, or nuclear systems. Multiphysics simulation has become an integral part of modern science and that improves the understanding of functionality, behavior, and even future condition in the considered system. Furthermore, it also saves time, money, and energy. The main goal of the simulation is to solve PDEs describing physical phenomena [1].

One of the most popular methods of solving PDEs is Finite Element (FE) method. Many commercial solutions, like e.g. COMSOL or ANSYS Multiphysics, and open source applications use FE method. One of the leading open source modules that implements FE method is libMesh FE library [1, 2]. Standard process of simulation starts with initialization of the system. Next, the mesh is

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_48, © Springer International Publishing Switzerland 2014

created and defined with specified parameters. Afterwards, a system of equations has to be created and defined with assemble function. The assemble function is responsible for providing pointers to mesh, system, degrees of freedom map, and FE object. In addition, the assemble function is responsible for filling matrices and vectors, called in FE terminology K^e and F^e respectively, with values. The values depend on cell-specific data like Jacobian quadrature weights, location of quadrature points on element, element shape function, and element shape function gradient. Finally, the created system is solved by one of the built-in or external solvers. In the entire process, the most time consuming part is matrix processing. Code analysis allows locating many spots where matrices are processed in multi-level loops. As it was shown in [1], the most time-consuming operation in FE-based approach is matrix processing that takes 45% of total simulation time. The most significant part from processing point of view is matrix multiplication function that can be found in DenseMatrixBase class. This function is referenced mostly by methods related to degrees of freedom. Fortunately, modern Graphic Processing Units (GPUs) are promising solution for FE-based multiphysics systems, which can reduce the matrix processing time [3].

In this paper, CUDA parallel programming model is used to implement two matrix multiplication algorithms. CUDA is a programming model developed by nVidia. It uses nVidia GPUs to perform general purpose computations on multiprocessors. The implemented parallel algorithms use CUDA threads to multiply matrices in more efficient and faster way. The reference point for the parallel algorithms is a classic single-thread matrix multiplication algorithm for Central Processing Unit (CPU) [4,5]. In this research, all considered CUDA algorithms and CPU algorithm are implemented to multiply two large size matrices. The implementation is done using discrete approach presented in [6]. Based on obtained results, the effectiveness of the presented CUDA algorithms is evaluated.

The rest of the paper is organized as follows. In Section 2 CUDA programming model is briefly presented and described. Section 3 contains description and design of three matrix multiplication algorithms. First of them is a classic version of algorithm designed for CPU, whereas two other algorithms are parallel CUDA algorithms executed by GPU. Experiments and results are presented and discussed in Section 4. Section 5 contains a summary of the paper.

2 CUDA

CUDA (Compute Unified Device Architecture) is a scalable parallel programming model and software environment developed by nVidia. Originally, CUDA was dedicated for nVidia GPUs, but recently the multicore and multiprocessor computers are also supported. The general idea behind parallel computing using GPU is to employ a parallel algorithm and many low-performance processing units in order to obtain better performance than high-performance single processor. General hardware structure of GPU is presented in Fig. 1 [7].

As it can be seen, GPU contains a number of independent multiprocessor units called Streaming Multiprocessors (SMs). Each SM has access to global

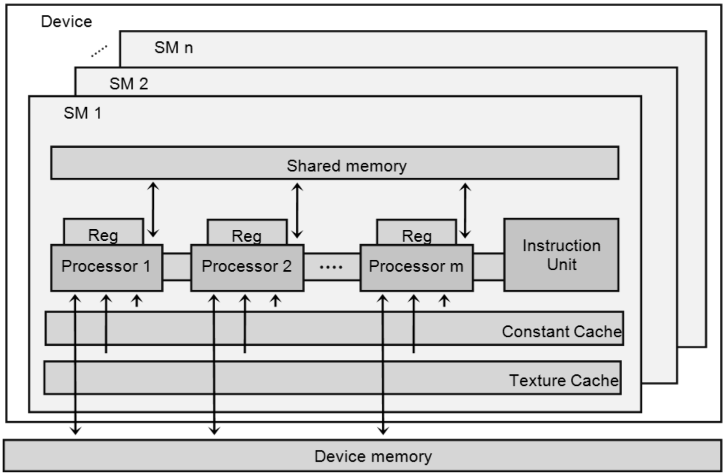


Fig. 1. GPU architecture [7]

memory. SM consists of processors connected with each other, instruction unit, shared memory (that is shared among all processors in single SM), texture cache, and constant cache. A set of registers is assigned to each processor. SM executes CUDA applications in parallel. Fig. 2 presents the logical structure of kernel [7,8].

Kernel is a single function that is executed by every thread that was assigned to execute the kernel. Threads are grouped in a 1D, 2D, or 3D grid. This grid of threads is called block. Blocks also form a grid; however, the dimension and size of the grid depend on the architecture of GPU. If SM contains 8 processors, then one dimension of grid of blocks is 8. Each block is mapped to a single SM. Therefore, block and threads share the resources available in SM. As long as the resources are available, threads are performed in parallel. The number of threads that can be executed in parallel without time sharing of resources on single SM is called warp. Each thread can be described by its identifier and block, which contains this thread. Based on these identifiers, kernel may perform different operations, although the code is the same for every thread [7].

Three types of memory are distinguished in CUDA programming model. First of them is global memory. It can be accessed by every thread from every block. This is the memory of the device. Access to this memory is fully synchronized. Global memory is slow and therefore it is not recommended to store frequently accessed data there. Second memory type is shared memory. This memory resides in SM and some part of the memory is assigned to a single block executed by SM. This part of memory can be accessed only by threads from the block to which memory was assigned. This memory is very fast so the most frequently used data, especially the one that is used by many threads in the same block, should be placed here. Third type of memory is local memory that can be used only by a thread that is assigned to this memory [7].

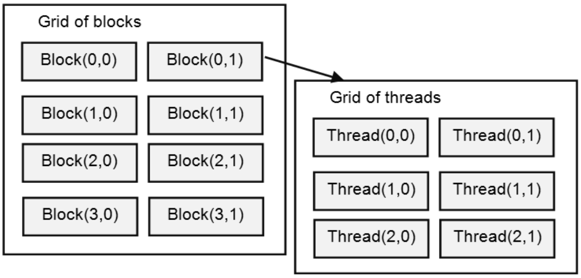


Fig. 2. Logical block/thread structure [8]

3 Matrix Multiplication Algorithms

In this section, three matrix multiplication algorithms are presented. First of them is designed for classic CPUs whereas, the next two are designed for CUDA enabled GPUs. Each of the algorithms extends the idea from previous one and adds new features. It is done intentionally and each next algorithm takes more advantages from GPU. Every algorithm is described, pseudo code is supplied; and advantages and disadvantages are presented.

3.1 CPU Multiplication Algorithm

First matrix multiplication algorithm is a classic approach designed for CPU. This method loops through all elements in result matrix C and calculates each single element. To calculate the value of single element, the algorithm loops through all elements of rows in matrix A and columns in matrix B; number of both rows and columns is specified. Pseudo code presented below describes the algorithm:

```
for $i = 0$ to $M$ do
    for $j = 0$ to $N$ do
        for $k = 0$ to $L$ do
            $C[i * L + j] += A[i * L + k] * B[j + k * L];$
        end
    end
end
```

Algorithm 1. CPU Multiplication Algorithm

2D matrices are presented as a vector consisting of rows from the original matrix. This approach simplifies the notation of high dimensional structures. Values M, N, and L are number of rows in matrix A, number of columns in matrix B, and number of rows and columns in matrix B and A, respectively.

Considered algorithm is easy to implement and to understand. However, the biggest disadvantage and advantage is that all computations are performed by one thread. Because only one thread is accessing the memory, the race condition does not appear and no synchronization is required. Unfortunately, this has significant impact on the effectiveness of the algorithm. For larger sized matrices, one thread is unable to obtain satisfactory execution time.

3.2 Basic CUDA Multiplication Algorithm

Second algorithm is a basic implementation of matrix multiplication algorithm in CUDA. In this approach, each element in result matrix C is calculated by a separate CUDA thread. Depending on the indices of thread and block that contains this thread, coordinates of element in C matrix are calculated. Thread gets the required elements from matrices A and B (that are stored in global memory) in a loop and calculate the element from matrix C [7,9]. Pseudo code clarifies the implementation of the algorithm:

```
$idA = (blockId.y * blockSize.y + threadId.y) * L$
$idB = blockId.x * blockSize.x + threadId.x$
$result = 0$
for $i = 0$ to $M$ do
    $result+ = A[idA] * V[idB]$
    $idA + +$
    $idB+ = N$
end
$idC = (blockId.y * blockSize.y + threadId.y) * L$
$idC+ = blockId.x * blockSize.x + threadId.x$
$C[idC] = result$
```

Algorithm 2. CPU Multiplication Algorithm

As in the previous algorithm, matrices are presented as vectors. M is the number of rows in matrix A, N is the number of columns in matrix B; and L is the number of rows and columns in matrix B and A, respectively. It may happen that no element in matrix C is assigned to some threads in a block. To avoid that situation, matrix size should be proportional to block size. Otherwise, some more conditional statements should be added to the algorithm, e.g. artificially setting zeros to elements that do not exist.

In this approach, process of calculating the values of elements in matrix C is distributed among number of threads and the process is performed in parallel. As a result, multiplying two matrices of significant size should be completed faster. Despite multithreaded execution, other advantages of the algorithm include simple design and implementation. However, the way of accessing the data strongly affects the effectiveness of the algorithm. Access to globally shared memory is slow mainly due to bus bandwidth and the need for synchronization. Therefore, the algorithm suffers from the number of needed accesses to matrices A, B, and C stored in global memory [10,11].

3.3 Tiled CUDA Multiplication Algorithm

The third multiplication algorithm is an extension of basic CUDA matrix multiplication algorithm. The third algorithm takes advantage of shared memory associated with every block of threads. This memory can be accessed only by threads that belong to a block with associated shared memory. Access to this memory is fast and efficient; however, the size of memory is small so it should be carefully used.

In the algorithm, the number of accesses to global memory is reduced by copying matrices A and B to shared memory. As it was mentioned earlier, shared memory has small capacity. Therefore, it is impossible and highly ineffective to copy entire input matrices. Thus only a small tile from both matrices A and B is copied to shared memory. The size of the tile is equal to the size of the block of threads. Each thread in a block copies one element from matrices A and B. It is important to let all of the threads to copy relevant data from input matrices before performing actual multiplication. Therefore, the second step of the algorithm is synchronization of threads. After that, when two tiles are copied to shared memory, multiplication is performed. Values of elements in matrix C are calculated by all threads that are contained in a single block [9, 10]. Pseudo code below presents the implementation of the algorithm:

```
$idA = (blockId.y * blockSize.y + threadId.y) * L + threadId.x$
$idB = blockId.x * blockSize.x + threadId.x + threadId.y * N$
$result = 0$
Create shared $tA$ and $tB$ of size $blockSize.y * blockSize.x$
for $i = 0$ to $L/blockSize.y$ do
    $tA[threadId.x + threadId.y * blockSize.x] = A[idA]$
    $tB[threadId.x + threadId.y * blockSize.x] = B[idB]$
    $idA += blockSize.x$
    $idB += blockSize.y * N$
    synchronize
    for $j = 0$ to $blockSize.y$ do
        $result = tA[j] * tB[j]$
    end
    synchronize
end
$idC = (blockId.y * blockSize.y + threadId.y) * L$
$idC += blockId.x * blockSize.x + threadId.x$
$C[idC] = result$
```

Algorithm 3. Tiled CUDA Multiplication Algorithm

Analogous to previous algorithms, matrices A, B, and C are given by vectors. Value N is the number of columns in matrix B; and L is the number of rows in matrix A and columns in matrix B. Similar to previous algorithm, appropriate size of the tile or additional condition statements should be added. However, despite these issues, the ratio of number of columns in A to number of columns in tile has to be equal to ratio of number of rows in B to number of rows in the tile.

The part of algorithm where data is copied from global memory to shared memory is especially interesting. It can be seen, that elements from matrices A and B are copied to horizontal vectors tA and tB. Therefore, tile taken from matrix B is transposed. This method of copying and storing data impacts the way vectors are multiplied. As it can be observed, element j from horizontal vector tA is multiplied with element j from horizontal vector tB. To adjust multiplication to classic algebraic rules, vector tB should be transposed. Nevertheless, this would cause unnecessary overhead and lower the effectiveness. Therefore, it is more reasonable to modify the method that multiplies vectors.

The final version of CUDA matrix multiplication algorithm uses most of the GPU features that allow obtaining high effectiveness. Values of elements in matrix C are calculated in parallel by several threads. Frequently accessed data is stored in fast shared memory. The main disadvantage of the algorithm is the design that is complicated and hard to understand. In addition, thread synchronization may impact the effectiveness, especially when the total number of threads exceeded the maximum number of threads that can be run in parallel. In this situation, threads share the resources and thread synchronization takes longer time [11].

4 Experiments

4.1 Experiments Environment

The experiments performed in this paper were carried out on nVidia Quadro 5000. According to device datasheet, graphic card supplies developers with one Quadro GPU with 352 CUDA cores clocked with 513 MHz each. The device offers 2.5 GB of GDDR5 (Graphic Double Data Rate v5) memory, 320 bit memory interface, and 120 Gbps memory bandwidth. GPU implemented in nVidia Quadro 5000 has 2.0 compute capability and it is called GF100. Generally, compute capability is a number that describes technology used in GPU. Capability 2.0 invokes following properties: number of SM is 32; maximum number of threads per block is 1024; maximum number of blocks in SM is 8; warp size is 32; maximum number of threads per SM is 1536; maximum number of 32 bit registers per thread is 63; maximum value of x and y dimension of block is 1024, and maximum z dimension is 64; maximum amount of shared memory per SM is 48 kB [8].

Since in Fermi architecture for each SM 32 CUDA cores are used, then Quadro 5000 offers 11 SMs. 8 blocks can be run on each SM; each SM can run up to 1536 threads. Therefore, the device supports up to 16896 threads grouped in 88 blocks. It is important to note, that developers have significant flexibility in defining quantity of blocks and quantity and grid structure that is formed by threads assigned to single block. Parameters mentioned above should not be exceeded; however, if they are exceeded, threads will not be executed in parallel and resources will be time-shared.

4.2 Plan of Experiments

Experiments presented in this paper consist two parts. The first part estimates the size of tile (parameter of Tiled CUDA algorithm) giving the best results. The second part of the experiments compares efficiency of the three presented algorithms.

To obtain the most promising size of tile, profiling of Tiled CUDA algorithm with different values of tile size was performed. Table 1 presents the structure of grids that were used. 2048×2048 matrices were used.

Table 1. Grids considered in experiments

Grid shape	Threads per block	Number of blocks	Number of threads
32×32	1024	16	16384
24×24	512	29	16704
16×16	256	66	16896
8×8	64	264	16896

In order to conduct second part of the experiments, six scenarios were prepared. Two square matrices were used for multiplication. They contained random generated 32 bit floating-point numbers in range between 0 and 10. The first of the scenarios contained 256×256 matrix size. Every next scenario doubled the matrix size. Therefore, matrix size of the last scenario was 8192×8192. To ensure a sufficient quality of obtained results, every scenario was repeated with 10 different matrices (the same matrix size but different values).

The grids presented in Table 1 satisfy all limitations of used GPU. Total number of threads is smaller or equal to maximum number of threads, thread grid matches size limitations, and the number of threads per block is smaller or equal to the maximum number of threads per block. In the tiled algorithm, each thread copies one 32 bit floating-point number, so each thread requires 4 B of memory. Since no more than 1536 threads are running in single SM, 48 kB limit of shared memory per SM is never reached.

4.3 Results

Table 2 presents the results of the first part of the experiment where the best size of tile was estimated. The tiled algorithm with different size of tiles was used to perform matrix multiplication.

As it can be seen, bigger the size of the tile, better are the results. Reduction of access to slow global memory in favor of fast shared memory allows obtaining data much faster (and therefore executing the multiplication faster). In basic

Table 2. Tiled CUDA algorithm performance in function of tiles size

Shape of grid	8×8	16×16	24×24	32×32
	0.90 s	0.57 s	0.42 s	0.32 s

CUDA matrix multiplication algorithm, each element in matrix A and B was accessed L times, where L is the number of rows and columns in matrix B and A respectively. Therefore for 2048×2048 matrix, first CUDA algorithm accesses global memory over 17 billion times just to get the data. In the tiled version of CUDA algorithm, each element in matrix A and B is accessed *L/blockSize.y* and *L/blockSize.x* times, respectively. Therefore, for the same matrix with tile size equal to 32×32, the algorithm accesses global memory only about 500 million times. Although increasing tile size is a promising way of increasing the performance, one has to be very careful not to exceed technology limits. As it was mentioned before, maximum number of threads per block is 32, therefore 32×32 is the highest possible size of tile.

Table 3 contains the profiling of each discussed algorithm.

Table 3. Algorithm effectiveness

Grid shape	Classic CPU algorithm	Basic CUDA algorithm	Tiled CUDA algorithm
256×256	0.12 s	0.23 s	0.18 s
512×512	0.90 s	0.34 s	0.19 s
1024×1024	5.64 s	0.87 s	0.22 s
2048×2048	32.55 s	1.53 s	0.31 s
4096×4096	554.6 s	2.32 s	0.45 s
8192×8192	11564.85 s	3.68 s	0.69 s

As it can be seen, classic CPU algorithm is more efficient only for the smallest size of the matrices. Even though, the difference in execution time among all the considered algorithms is not large for the smallest matrices. For all other considered matrices, CUDA algorithms outperform CPU algorithm. It can be observed that increasing the size of matrices causes rapid increase in execution time for CPU algorithm, whereas execution time of CUDA algorithms presents linear and almost flat execution time increase. It also can be seen that the tile version of CUDA scheme was almost 5 times better than the basic CUDA algorithm for all matrix sizes.

5 Conclusions

In this paper, two CUDA matrix multiplication algorithms were compared to the classic algorithm for CPU. First CUDA algorithm was the simplest parallel algorithm where each result matrix element was calculated by separate thread. In the second CUDA algorithm, frequency of accessing global memory was significantly reduced by using shared memory. Only for the smallest size of examined matrices, CPU algorithm was more effective than CUDA algorithms. For all other sizes, both CUDA parallel algorithms significantly outperformed classic CPU algorithm. The greater matrix size, the difference between execution time of CUDA algorithms and CPU algorithm is higher. It was also shown that extensive use of shared memory within a block gives far better result than using only global memory.

This paper opens wide spectrum of possible further work. First of all new parallel multiplication algorithms may be designed and tested. Examining different structure of thread grids or even a dynamic structure whose shape depends on input matrices is also a reasonable further research direction. Finally, presented Tiled CUDA algorithm may be implemented in a multiphysics system, FE library or FE-based solver.

References

1. Krol, D., Zydek, D.: Solving PDEs in Modern Multiphysics Simulation Software. In: 2013 IEEE International Conference on Electro/Information Technology (EIT 2013), pp. 1–6 (2013)
2. libMesh webpage (2013), `http://libmesh.sourceforge.net/examples.php`
3. Liu, B., Zydek, D., Selvaraj, H., Gewali, L.: Accelerating High Performance Computing Applications Using CPUs, GPUs, Hybrid CPU/GPU, and FPGAs. In: 2012 13th Inter. Conf. on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2012), pp. 337–342 (2012), doi:10.1109/PDCAT.2012.34
4. Zydek, D., Selvaraj, H., Gewali, L.: Synthesis of Processor Allocator for Torus-Based Chip MultiProcessors. In: 7th Inter. Conf. on Information Technology: New Generations (ITNG 2010), pp. 13–18 (2010), doi:10.1109/ITNG.2010.145
5. Zydek, D., Chmaj, G., Chiu, S.: Modeling Computational Limitations in H-Phy and Overlay-NoC Architectures. The Journal of Supercomputing (2013), doi:10.1007/s11227-013-0932-9
6. Chmaj, G., Zydek, D.: Software Development Approach for Discrete Simulators. In: 21st International Conference on Systems Engineering (ICSEng 2011), pp. 273–278 (2011), doi:10.1109/ICSEng.2011.56
7. Nvidia: CUDA Programming Guide 2.0. Technical report, Nvidia (2009)
8. nVidia webpage (2013), `http://developer.nvidia.com/object/cuda.html`
9. Ryoo, S., et al.: Optimization Principles and Application Performance Evaluation of a Multithreaded GPU using CUDA. In: 13th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, pp. 73–82 (2008)
10. Cecilia, J.M., et al.: The GPU on the simulation of cellular computing models. Soft Computing 16(2), 231–246 (2012)
11. Fatahalian, K., Sugerman, J., Hanrahan, P.: Understanding the efficiency of GPU algorithms for matrix-matrix multiplication. In: ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware, pp. 133–137 (2004)

Tracker-Node Model for Energy Consumption in Reconfigurable Processing Systems

Grzegorz Chmaj, Henry Selvaraj, and Laxmi Gewali

Howard R. Hughes College of Engineering, University of Nevada Las Vegas, USA
{Grzegorz.Chmaj,Henry.Selvaraj,Laxmi.Gewali}@unlv.edu

Abstract. In this paper, we present an energy dissipation model for reconfigurable systems in which FPGAs have the property of online reprogramming. The proposed system contains regular nodes and one control node. Each regular node contains both CPU - capable of software processing, and FPGA unit which after being programmed with bitstream serves as the hardware processing parts. Nodes are connected in some structure and the connections form the transport layer. The system is capable of processing tasks in a distributed manner and communication, control and processing parts are taken into consideration in the energy equations. The model has also been used for algorithms that formed the complete system that is used for experimentation.

Keywords: reconfigurable processing, FPGA, distributed computing.

1 Introduction

Reconfigurable Systems (RS) are current trend in distributed processing. They are built of nodes that allow partial hardware reconfiguration. This gives them more flexibility over non-reconfigurable structures in processing needs, such as higher efficiency and lower power consumption. Compared to Application-Specific Integrated Circuits (ASIC), RS offers short reconfiguration time, the ability of multiple reconfiguration and low price of the reconfigurable unit. Nodes in RS are connected with each other using the interconnection structure (IS) of given topology, which also impacts the energy used to compute the given task. Contribution of this paper is to introduce the preliminary modeling of reconfigurable system for tracker-node architecture, which is the base for further research. We show the complete model of energy dissipation, two algorithms based on the presented model and experimentation results.

2 Literature Overview

Distributed processing structures are used to lower the financial costs of process intensive tasks. Industry applications widely use grids, formed mostly by groups of institutions and later sharing the grid's resources [3]. Grids however require the

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_49, © Springer International Publishing Switzerland 2014

financial investments, this led to the foundation of private distributed computation networks, such as SETI [8], where the power of personal computers is used. The processing power currently is a valuable asset, so the investments to the dedicated computational structures are worth considering [10]. The most efficient approach is to use ASICs [9], however they are designed to suit the specific tasks, what makes them hard to use for other types of tasks. FPGAs can be repeatedly reprogrammed, making the system adjustable to changing needs [1], [2]. This makes the reconfigurable systems an interesting topic of research, gaining the attention of many research institutions. They are considered both as on-chip systems, called RSoC [4] and large scale structures [5]. Wide spectrum of reconfigurable systems applications is a subject of research, e.g. image processing [6], unmanned aerial vehicles [7].

3 Tracker-Node Structure

We propose the structure of reconfigurable system using tracker-node approach for its operation. Tracker is a special node that handles the control over the nodes. The general system scheme is shown in Fig. 1:

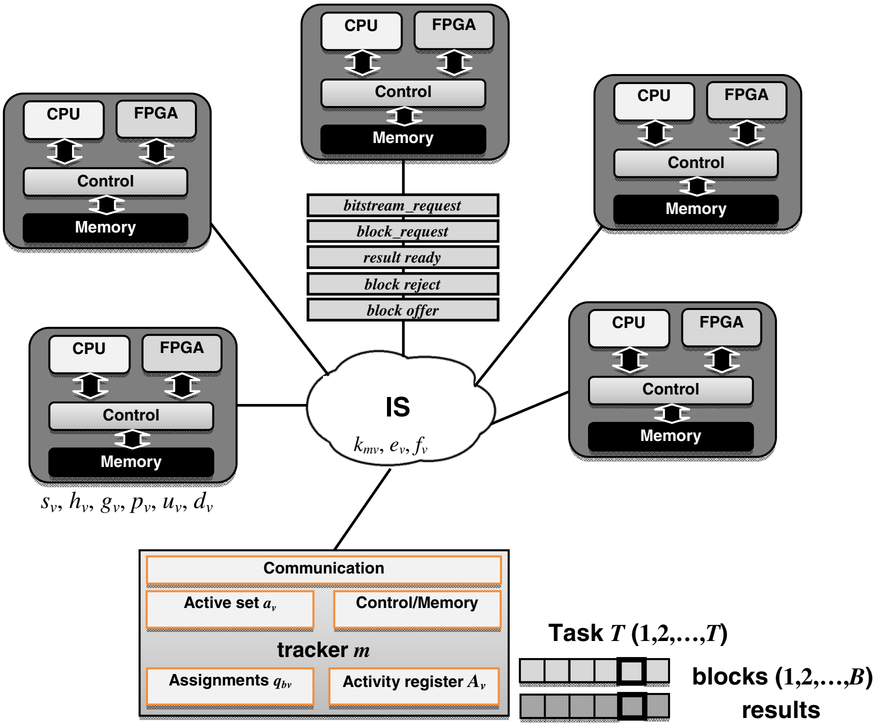


Fig. 1. General system diagram

Nodes and tracker are connected using IS, which can be defined to reflect any topology.

3.1 System General

V nodes are present in the system (including tracker):

$$v, w = 1, 2, \ldots, V\,. \tag{1}$$

Processing (input) task is divided into *B* blocks of the same size:

$$b = 1, 2, \ldots, B\,. \tag{2}$$

Processing of block b yields the result r (same id):

$$r = 1, 2, \ldots, B\,. \tag{3}$$

Operating timespan is divided into *T* undividable slots:

$$t = 1, 2, \ldots, T\,. \tag{4}$$

Block b is computed at the node v at time t (index):

$$x_{bvt} = 1 \text{ (0 otherwise)}\,. \tag{5}$$

Tracker is the special node in the system, denoted as *m* (6)

Reconfigurable system is used to process the given task, which is split into chunks (blocks) for the purpose of processing (this also means that the task must be divisible) (2). For the sake of simplicity, this paper considers tasks divided into uniform blocks (i.e. having the same size). The division operation occurs at the special node called tracker, which also performs the role of coordinator. Blocks are then sent to nodes for processing. Once the block is processed, the result of computation has the form of result blocks *r* (3).

The system contains *V* nodes of the same parameters. Each node contains processing unit capable of performing software functions, and the reprogrammable FPGA unit – with the capability of online reconfiguration. The online reconfiguration allows FPGA to be programmed during normal system operation with the received bitstream. The programing part is performed by the control unit, which is a part of the node. The system operates in real time, for better description of the system and algorithms, and for precise properties description, we consider the timespan divided into time slots (4). This brings the ability to fully express the moment of each event.

3.2 Node

Energy used on node *v* for computation by software:

$$s_v = \text{const}\,. \tag{7}$$

Energy used on node v for computation by hardware:

$$h_v = \text{const} . \tag{8}$$

Node v has the bitstream for hardware computing (index):

$$g_v = 1 \text{ (0 otherwise)} . \tag{9}$$

Node v has limited computation capability:

$$p_v = \text{const} . \tag{10}$$

Each node can decide to fetch the bitstream from the tracker node. Once the bitstream is fetched, the node gathers the ability to perform the hardware computation (9). Both hardware and software computations involve the given amount of energy (7), (8). In this paper we assume that $h_v < s_v$, thus the benefit of using hardware computing is that the amount of energy emitted is smaller.

3.3 IS Properties

Energy emitted by sending the block b from tracker to node v:

$$k_{mv} = \text{const} . \tag{11}$$

Energy emitted by sending the result r from node v to tracker:

$$k_{mv} = \text{const} . \tag{12}$$

Block b is sent from node w to node v at the time t (index):

$$y_{bwvt} = 1 \text{ (0 otherwise)} . \tag{13}$$

Bitstream is sent from the tracker to node v at the time t (index):

$$z_{wvt} = 1 \text{ (0 otherwise)}, w=m . \tag{14}$$

The cost of sending the bitstream from tracker to node v:

$$e_v = \text{const} . \tag{15}$$

Time required for fetching the bitstream from tracker to node v:

$$f_v = \text{const} . \tag{16}$$

Time required for transfer of block/result between node v and tracker:

$$j_v = \text{const} . \tag{17}$$

The IS operation also involves electrical energy. In this paper, we assume that the energy used for sending the block from tracker to node, equals the energy required for sending the result back from this node to the tracker (11), (12). The moment of transfer of the block or the bitstream is determined by (13) and (14). Similar relations are formulated for bitstreams (14), (15). (16) and (17) express the time, which is required to transfer blocks, results and bitstreams between the given node and the tracker.

3.4 Constraints

Certain block b is processed only at one node:

$$\sum_v \sum_t x_{bvt} = 1 \quad b = 1, 2, \ldots, B\ . \tag{18}$$

Each node v has limited computational capabilities:

$$\sum_b \sum_t x_{bvt} \le p_v \quad v = 1, 2, \ldots, V\ . \tag{19}$$

Each node w has limited upload capabilities:

$$\sum_b \sum_v y_{bwvt} \le u_w \quad w = 1, 2, \ldots, V \;\; t = 1, 2, \ldots, T\ . \tag{20}$$

Each node v has limited download capabilities:

$$\sum_b \sum_w y_{bwvt} \le d_v \quad v = 1, 2, \ldots, V \;\; t = 1, 2, \ldots, T\ . \tag{21}$$

All results have to be sent back to the tracker:

$$\sum_r \sum_v y_{rwvt} = 1 \quad v{=}m\ . \tag{22}$$

Computation of block b at the node v can finish when b is done fetching:

$$\sum_{t=1..q} y_{bwvt} + \sum_{t=q+1...T} x_{bvt} = 2 \quad w{=}m,\ q \ge j_v\ . \tag{23}$$

Constraints define the assumptions for the system. Each node has the limited computational capabilities (19), (10) that determines the amount of data it can process in given timespan. Regarding node's communications capabilities, upload (20) and download (21) speeds are defined. The processing is considered as finished when all *B* result blocks are collected at the tracker node (22). For this paper, we assume that each block will be assigned for processing to only one node (18), and that the computation of block *b* can start when *b* is fully downloaded – no processing of partially fetched block is allowed (23). The constraint (23) also assures that the block fetched by node *v*, will be processed by this node. The goal of the system is to process the task, divided into *B* blocks and collect the results at the tracker node. This will yield the following energy emission components:

Fetching bitstreams:

$$\sum_v \sum_t z_{mvt} g_v e_v\ . \tag{24}$$

Fetching blocks:

$$\sum_b \sum_v \sum_t x_{bv}\, y_{bmvt}\, k_{mv}\ . \tag{25}$$

Performing the computations:

$$\sum_b \sum_v \sum_t (g_v h_v + (1-g_v) s_v) x_{bvt} \, . \tag{26}$$

Results return:

$$\sum_r \sum_v \sum_t x_{rv} \, y_{rvmt} \, k_{mv} \, . \tag{27}$$

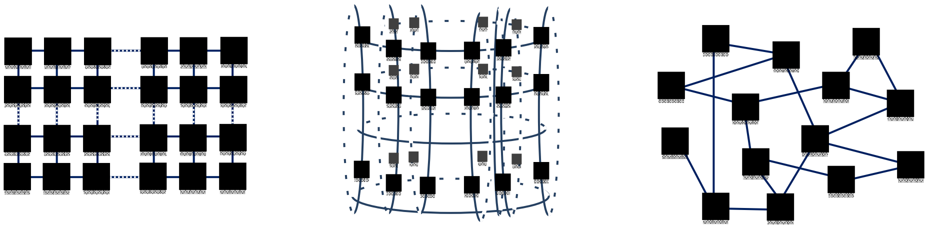
The overall energy consumption will be the sum of the components:

$$E = \sum_v \sum_t z_{mvt} g_v e_v + \sum_b \sum_v \sum_t x_{bv} \, y_{bmvt} \, k_{mv} + \sum_r \sum_v \sum_t x_{rv} \, y_{rvmt} \, k_{mv} + \sum_b \sum_v \sum_t (g_v h_v + (1-g_v) s_v) x_{bvt} \, .$$

According to the node algorithm, nodes can decide to fetch the bitstream or perform the computations based on the software. If a node will decide to fetch the bitstream (9), it will be done with the energy e_v (15). This energy characterizes the network relation between tracker and the fetching node, and can be different for each node. If a node is fetching the bitstream (14) during time t, we assume that this process ends in time slot $t+f_v$ (according to (14) and (16)). To perform the computation, a node fetches the block from the tracker, generating the cost (25). This transfer is indicated in (25) to be started in time slot t, and will end in time slot $t+j_v$ (17). To be able to perform the block processing, a node has to finish block fetching – there is no possibility to process partially fetched block (23). (26) describes the energy emitted in during the process of computation: already fetched block b is either processed using software $(1-g_v)s_v x_{bvt}$ or hardware $g_v h_v x_{bvt}$. Variable g_v assures that either one of these two costs will be produced. After block computation, when the result r is produced, it is sent back to the tracker.

3.5 Communications

IS is the vital part of the processing system. We propose the communication layer based on message-exchange protocol. The following messages are being used in our system: *bitstream_request*, *block_request*, *result_ready*, *block_reject*, *block_offer*. The IS structure is defined using values of k_{mv} (11), (12) – determining the cost of blocks and results transfer, and bitstream sending e_v (15). The energy for node may be interpreted as the distance from the node to the tracker node. This way k_{mv} and e_v can describe the physical structure – three structures considered in this paper are shown in Fig. 2. Mesh is the regular interconnection network, where nodes form the matrix (Fig. 2 a). The torus is created based on mesh – the connections do not end on the boundary nodes, but form a connection to the overlapping node (Fig 2. b)). The third considered IS is freely unstructured, where the lengths of inter-node connections do not follow any specific rule (Fig. 2. c)). The IS structure also determines the timing relations in the system. In this paper, we simplified them by using two variables: f_v determining the time required for fetching the bitstream to node, and j_v – determining the block/result transfer time (we consider the link to be symmetric). Constraints (20) and (21) are also related to timing, as they determine the transfer speeds – each node has limited download and upload capability.



a) mesh b) torus c) unstructured

Fig. 2. Examples of IS structures

Tracker node tracks the nodes activity by using active nodes set (28) and action register (29). The first one contains all nodes known to the tracker (not all nodes are known to the tracker when system starts the operation), the latter indicates the status of the node, although this information may become outdated.

Node is active (index): a_v=1 (0 otherwise) (28)

Node action: A_v={*idle, fetch, send, processing*} (29)

Block *b* is assigned to node *v*: q_{bv}=1 (0 otherwise) (30)

Block is processed (index): r_b=1 (0 otherwise) (31)

The operation of system elements are described by algorithms. All nodes operate under the same algorithm, the tracker node has its own specific algorithm. Algorithms are using request-respond architecture. Nodes decide whether to fetch the bitstream, or directly start requesting blocks. To fetch the block, node sends the request to the tracker, which responds with the block and expects the return of the result. Tracker algorithm serves the requests from nodes and keeps track of results (can also request the node which it considers as idle) and combines the final result.

4 Experimentation Results

The system described above was implemented as a software simulator. It takes in several parameters, which characterize the experiment: IS topology, the energy required for transferring data and bitstreams among the nodes, etc.

The first experiment shows the impact of using hardware processing, compared to software processing. IS structure used in this case is unstructured (Fig. 2.c)). The IS contains 50 nodes, the average energy consumption per block processing is 14.2mW for software processing and 5.2mW for hardware processing. Energy consumption for data transportation in IS ranges between 3-49mW per uniform data block. The bitstream fetch consumes an average of 10.2mW. Figure 3 shows the same IS structure in three cases: all nodes perform hardware processing (allHw), all nodes perform software processing (allSw) and mixed – where only some nodes process task using hardware (24 of them) and the rest of them use the software approach. These three cases are used for various task sizes *T*. For *T*=200 – the differences between allHw, allSw and mixHS – are small: allHw required 15.3% less energy than allSw, 13.5% for mixHS. For small task sizes, the energy used to fetch bitstreams negatively influenced the overall cost, as only small number of blocks are processed by hardware. For *T*=500, energy saving increases (comparing to allSw: 18,8% less for allHw and 15.6% less for mixHS). This trend continues for *T*=1500 and *T*=5000 (19.8% and 20.3% allSw to allHw, and 16.2% and 16.5% allSw to mixHS respectively).

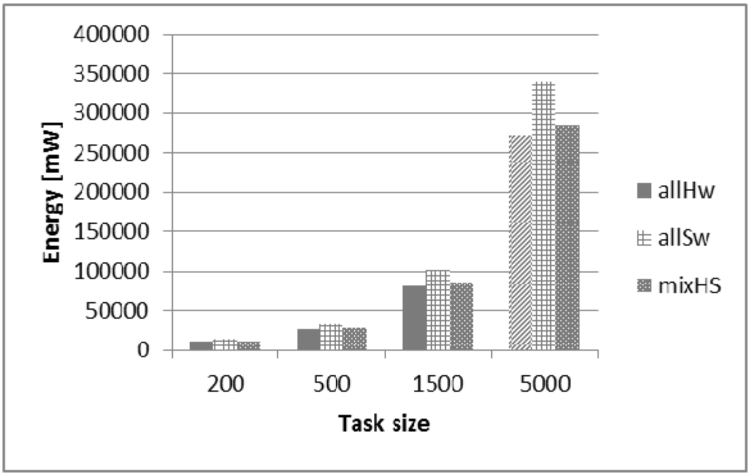


Fig. 3. Three processing cases

The relation between h_v and s_v is defined for the whole system as the ratio $R=\sum_v \frac{h_v}{s_v V}$. Fig.4 presents the experiment for three *R* values: *R*=6.26, *R*=28.4 and *R*=64.1. The mixHS case was used. Experiments show that small task sizes require more or less the same amount of energy. The advantage of the proper *R* ratio becomes more visible as the processing load increases. The average energy consumption per block (including transfer, computation, result return and share in bitstream processing) was equal to 58.4mW. The energy saved by using *R*=6.26 instead of *R*=64.1 would allow computation of additional 3154 data blocks using the same system resources. For *T*=1500 the energy saving, compared to *R*=6.26 was 92W and 184W for *R*=28.4 and *R*=64.1 respectively. Experiments also show, that the relation between energy saving and *R* ratio is not linear – other aspects such as energy required for bitstream fetch, communication time and mutual relations between nodes also impact the final energy dissipation amount. For too small *R* values, system becomes inefficient for hardware processing – the research about finding the proper *R* ratio for wide range of input conditions is the part of our current and future work.

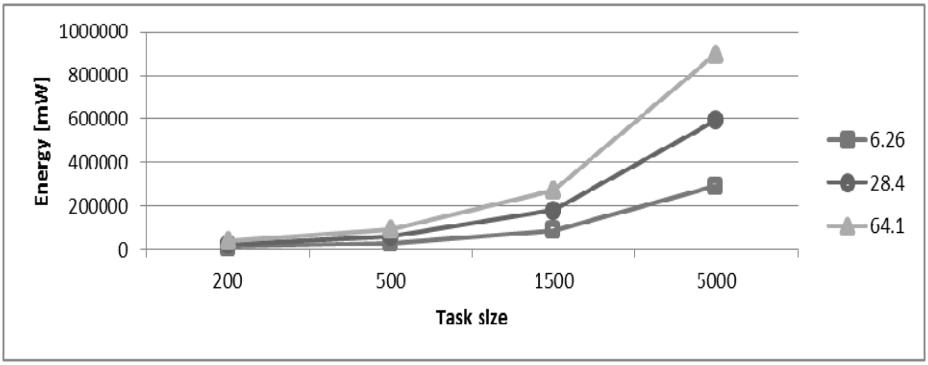


Fig. 4. The relation between *m* ratio and energy dissipation for various tasks

The IS structure impacts the operational energy – the part responsible for data transfer. Our experiments show that the mesh and torus cases resulted in very similar energy dissipation (below 5% of difference). Unstructured IS demonstrated its advantage for larger tasks (13% less energy emitted for task size *T*=500 and larger).

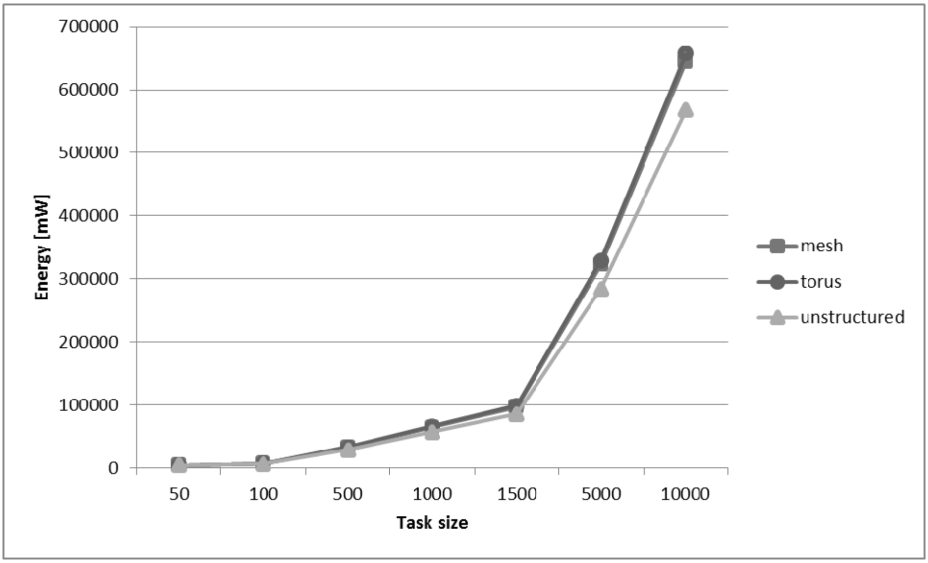


Fig. 5. Energy dissipation for three structures

5 Conclusion

Reconfigurable systems provide many possibilities to act as flexible structures, which in turn save the overall energy used. The model presented in this paper is being extended in our current research in order to be able to handle more complex processing tasks. The experimentation results using two presented algorithms show that hardware processing can lower the energy used for computation, but the system configuration is not straightforward for all cases. Our further research concentrates on using many bitstreams, task types and nodes with multiple reconfiguration units.

References

1. Gokhale, M.B., Graham, P.S.: Reconfigurable Computing, Accelerating Computation with Field-Programmable Gate Arrays. Springer (2005)
2. Hauck, S., Dehon, A.: Reconfigurable computing – The theory and practice of FPGA-based computing. Morgan Kaufmann/Elsevier (2008)
3. Mahajan, S., Shah, S.: Distributed Computing. Oxford University Press (2010)
4. Samara, S., Schomaker, G.: Self-adaptive OS Service Model in Relaxed Resource Distributed Reconfigurable System on Chip (RSoC). Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, 1–8 (2009)
5. Nadeem, M.F., Ostadzadeh, S.A., Nadeem, M., Wong, S., Bertels, K.: A Simulation Framework for Reconfigurable Processors in Large-scale Distributed Systems. In: 2011 International Conference on Parallel Processing Workshops, pp. 352–360 (2011)
6. Salvador, R., Otero, A., Mora, J., De la Torre, E., Riesgo, T., Sekanina, L.: Self-reconfigurable Evolvable Hardware System for Adaptive Image Processing. IEEE Transactions on Computers (2013)
7. Jasiunas, M., Kearney, D., Bowyer, R.: Connectivity, Resource Integration, and High Performance Reconfigurable Computing for Autonomous UAVs. In: IEEE Aerospace Conference, pp. 1–8 (2005)
8. Korpela, E., Werthimer, D., Anderson, D., Cobb, J., Lebofsky, M.: SETI@home-massively distributed computing for SETI. Computing in Science & Engineering 3(1), 78–83 (2001)
9. Ballinger, N.: ASIC technical training-the challenges and opportunities. In: Proceedings of Second Annual IEEE ASIC Seminar and Exhibit, pp. P14-4/1-4 (1989)
10. Zydek, D., Selvaraj, H., Gewali, L.: Synthesis of Processor Allocator for Torus-Based Chip MultiProcessors. In: Proceedings of 7th International Conference on Information Technology: New Generations (ITNG 2010), pp. 13–18. IEEE Computer Society Press (2010), doi:10.1109/ITNG.2010.145

A Nonlinear Analysis of Vibration Properties of Intracranial Saccular Aneurysms

Kasper Tomczak and Roman Kaszyński

Faculty of Electrical Engineering, West Pomeranian University of Technology,
ul. Sikorskiego 37, 70-313 Szczecin, Poland

Abstract. In this paper, a nonlinear analysis of vibration properties of intracranial saccural aneurysms is presented. Intracranial saccural aneurysms have been clinically observed to emit a sound (a bruit) on each heartbeat. Our hypothesis about the reason of the sound is that the bruit is caused by resonance of aneurysm's walls. We apply a nonlinear analysis of the vibration properties to show that the resonance of aneurysms is possible. At the end of the paper an experiment *in silico* is carried out and conclusions are drawn.

Keywords: nonlinear analysis, vibration properties, aneurysm, cardiovascular disease.

1 Introduction

The cardiovascular diseases, wide type of cancers and degenerations in walls of arteries are one of the main causes of death. From these group of wall degeneration we can specify intracranial aneurysm (*aneurysma intracraniale, sacculare*) [1], which can be defined as a pathological bifurcation or over-stretched artery wall. Usually, they are formed in weakened places of artery and take a form of *sac* about 5 to 10 mm of diameter up to 25 mm in some cases. According to the literature [9], [19] the occurrence of this disease is in range of 0,5-2% or even up to 5% of the world population. Typically, aneurysms are detected after the rupture, that is, after the subarachnoid hemorrhage (SAH) whose effect are mostly fatal. Therefore, there is a constant need to develop new theoretical studies based on mathematical modelling which can support clinical investigations.

Methods of mechanical modelling of the aneurysm, which aim at identifying the critical parameters that cause the rupture, do not respond clearly to the question about the reason of the aneurysm rupture [6]. Therefore, it seems necessary to propose a different approach to this topic. Very interesting and appropriate is to examine whether in the artery - aneurysm system may be at resonance or not [20]. Works about resonances [5], [11], [13], [15], [16], [17] suggest that such phenomenon may occur in aneurysm. This could explain why aneurysms rupture not only during a sudden increase in blood pressure, but also during sleep. Locksley et al. [10] analyzed 2288 cases of aneurysms with the known activity or event associated with SAH and showed that cracking occurs frequently during sleep (36%). This statistic shows that aneurysm rupture can occur at normal

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_50,

frequency and blood pressure. Jain [8] wrote that pulsating flow can cause rupture of the aneurysm by the excitation of the resonant frequencies in the sac. He claimed that the sounds heard in the aneurysm confirmed the existence of this phenomenon. Ferguson [4] wrote that the noises recorded in a state of sleep occur in 12 out of 19 cases studied. Recording device in these situations has been applied to the outer surface of the surgically exposed blood vessels and aneurysms. In cases where the noises are not registered, the flow in aneurysm was limited. The frequency noise was in the range of 330 to 590Hz with an average 440Hz [4], [7]. In more recent works [2], [15] measured frequency range was 200 to 800 Hz and from 100 to 1000 Hz [14]. There are three main hypotheses explaining the existence of the noise:

- the existence of turbulence in the aneurysm's sac,
- local blood vortexes in the aneurysm,
- self-excited oscillation.

It seems that the combination of pulsating blood flow in the parent vessels and the size of the aneurysm and its entry at the root rises to vibrations that can be heard as a sound, or noise.

In this paper, we are interested in frequency analysis of biomedical systems, which are intracranial aneurysms. Due to the fact that experiments can damage patients cranial and brains of patients, frequency analysis will be undertaken using mathematical modelling and tools used in the field of systems theory and automatic control engineering. This is a well-known procedure in the modelling of biomedical objects [18], [21]. The obtained resonant frequencies can be compared with audible frequencies, which will identify aneurysms that are particularly vulnerable to cracking. In addition, the gain which occurs in the resonance gives the answer how big is the risk of rupture, when an aneurysm is in resonance.

The frequency analysis of the aneurysm-artery system is used as a model in which the aneurysm is treated as a spherical membrane filled with fluid (blood). Originally, this approach was used to determine the vibration of cranial filled cerebral spinal fluid [22]. However, this model can also be used to describe a far smaller structures which contain fluid at any density. In this case it is an aneurysm filled with blood.

The work consists of the following parts. Section 2 describes the model of the aneurysm as a liquid-filled spherical membrane. In section 3, simulations are carried out. Section 4 draws conclusions.

2 Aneurysm as a Liquid-Filled Spherical Membrane

2.1 Introduction

Aneurysm may be considered as a fluid-filled spherical membrane. Since blood is an incompressible fluid, so it has to be taken into the considerations for such arrangements. Young and Egin in their works [3], [22] gave an equation for the vibrations for any type of membrane filled with an incompressible fluid, but their

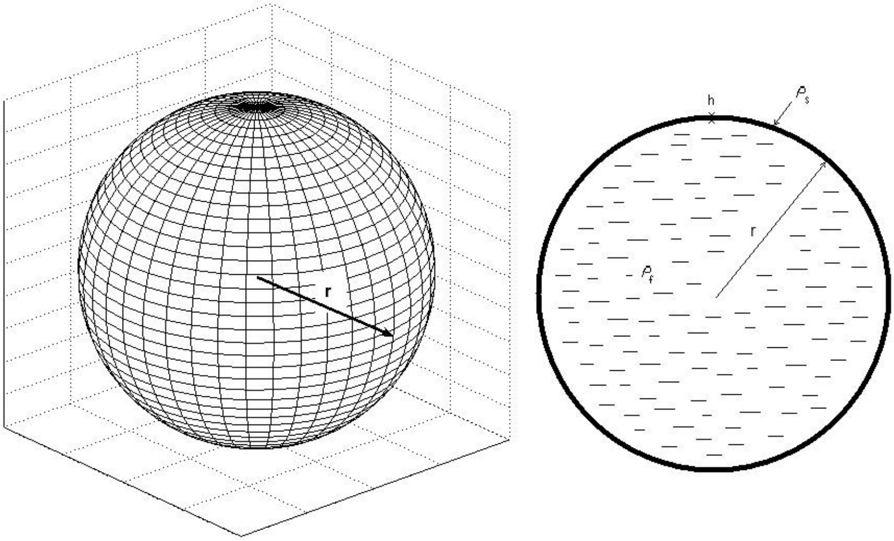


Fig. 1. Fluid-filled sphere

deliberations were used to model the cranial filled with cerebral spinal fluid. This section describes the use of this equation for cerebral aneurysm and allows to calculate its natural frequency.

2.2 Mathematical Model

Now consider the aneurysm as a spherical membrane filled with an incompressible fluid (blood) and $n > 0$ vibration modes, radius r, wall thickness h, Young modulus E_s, density ρ_s, the speed of the pressure wave in the shell $c_s^* = \sqrt{\frac{E_s}{\rho_s(1-\upsilon^2)}}$ and Poisson's coefficient υ. Furthermore, by ω denote the radial frequency of natural membranes, while for ρ_f and $c_f = \sqrt{\frac{B}{\rho_f}}$ - respectively - the density and wave velocity of the fluid pressure (B - weight modulus of the liquid). Then the equation for the free vibrations of the membrane takes the following form [22]:

$$\beta^4(1+\tau) - \beta^2\big(1 + 3\upsilon + \lambda_n - (1 - \upsilon - \lambda_n)\tau\big) - (1-\upsilon^2)(2-\lambda_n) = 0 \qquad (1)$$

where:
$\beta = \frac{\omega r}{c_s^*} = \omega r\sqrt{\frac{\rho_s(1-\upsilon^2)}{E_s}}$;
$\lambda_n = n(n+1)$;
$\tau = \frac{\rho_f}{\rho_s}\ \frac{r}{h}\ \frac{1}{n}$.

For further consideration it is convenient to introduce the dimensionless parameter

$$\varOmega = \omega r \sqrt{\frac{4\pi}{3}\frac{r}{h}\frac{\rho_f + 3(\frac{h}{r})\rho_s}{E_s}} \tag{2}$$

which for thin layers is approximately $\varOmega \approx \omega\sqrt{\frac{mass}{hE_s}}$, where $mass$ is the mass of the layer and the liquid contained therein.

Let us now determine β, depending on the dimensionless parameter

$$\varOmega = \omega r \sqrt{\frac{4\pi}{3}\frac{r}{h}\frac{\rho_f + 3(\frac{h}{r})\rho_s}{E_s}} \Longrightarrow \omega r = \varOmega\sqrt{\frac{3}{4\pi}\frac{h}{r}\frac{E_s}{\rho_f + 3(\frac{h}{r})\rho_s}}.$$

Substituting this expression into the formula for β, after the simplifications and transformations one obtains

$$\beta = \varOmega\sqrt{\frac{3}{4\pi}(1-\upsilon^2)\frac{1}{(\frac{r}{h})\frac{\rho_f}{\rho_s} + 3}}, \tag{3}$$

where $\frac{\rho_f}{\rho_s} = n\tau$. Finally, the size of β expresses by the dimensionless parameter is:

$$\beta = \varOmega\sqrt{\frac{3}{4\pi}\frac{1-\upsilon^2}{n\tau + 3}}. \tag{4}$$

Inserting (4) to (1) yields

$$\varOmega^4\frac{(1-\upsilon^2)^2}{(n\tau+3)^2}\Big(\frac{3}{4\pi}\Big)^2(1+\tau) - \varOmega^2\frac{1-\upsilon^2}{n\tau+3}\Big(\frac{3}{4\pi}\Big)(1+3\upsilon+\lambda_n-(1-\upsilon-\lambda_n)\tau)-$$

$$-(1-\upsilon^2)(2-\lambda_n) = 0. \tag{5}$$

After substitution:

$$a = \frac{(1-\upsilon^2)^2}{(n\tau+3)^2}\Big(\frac{3}{4\pi}\Big)^2(1+\tau),$$

$$b = \frac{1-\upsilon^2}{n\tau+3}\Big(\frac{3}{4\pi}\Big)(1+3\upsilon+\lambda_n-(1-\upsilon-\lambda_n)\tau),$$

$$c = (1-\upsilon^2)(2-\lambda_n),$$

$$X = \varOmega^2,$$

equation (5) takes the following form:

$$aX^2 - bX - c = 0. \tag{6}$$

The solutions of the equation (6) are:

$$\varOmega_{1,2} = \pm\sqrt{\frac{b \pm \sqrt{b^2+4ac}}{2a}},$$

It has to be noted that negative values cannot be observed in the real world, so we are interested only in the positive ones

$$\Omega = \sqrt{\frac{4\pi}{3}(n\tau + 3)\frac{(1 + 3\upsilon + \lambda_n - (1 - \upsilon - \lambda_n)\tau)\pm}{2(1 - \upsilon^2)(1 + \tau)}}$$

$$\overline{\frac{\pm\sqrt{(1 + 3\upsilon + \lambda_n - (1 - \upsilon - \lambda_n)\tau)^2 + 4(1 + \tau)(1 - \upsilon^2)(2 - \lambda_n)}}{2(1 - \upsilon^2)(1 + \tau)}}. \tag{7}$$

Chart of the dimensionless parameter as a function of τ, described by the relation (7), is shown in Figure 2.

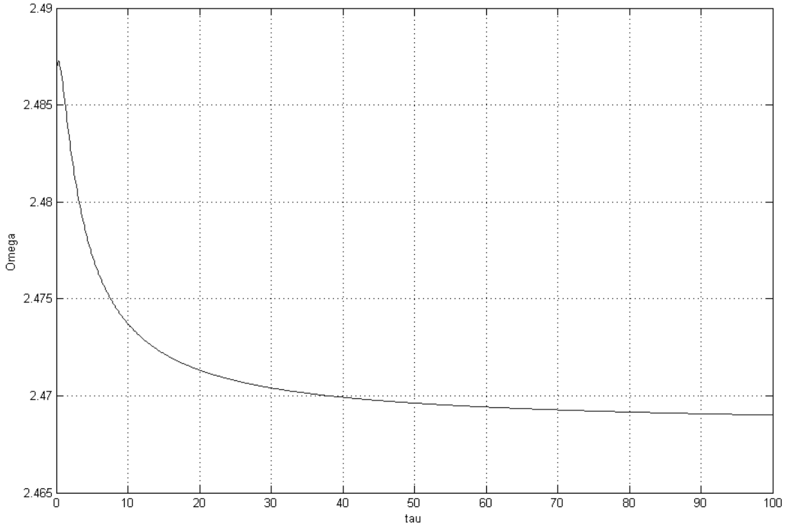


Fig. 2. Dependence $\Omega = \sqrt{\frac{b - \sqrt{b^2 + 4ac}}{2a}}$ from τ, for $\upsilon = 0.5$ and $n = 2$

2.3 Final Solution

Let us now determine natural frequencies from equation (7):

$$\omega = \Omega\frac{1}{r}\sqrt{\frac{3}{4\pi}\frac{h}{r}\frac{E_s}{\rho_f + 3\frac{h}{r}\rho_s}}. \tag{8}$$

In order to determine the frequency in Hertz, it is important to divide the ω by 2π which yields final solution:

$$f = \frac{\omega}{2\pi}. \tag{9}$$

3 Experiment

3.1 Results

Using the sphere to describe aneurysm, τ cannot reach high values. Therefore, one must link parameters from empirical data observed during the study of aneurysms. The values of the parameters required to determine the natural frequencies take the values shown in the table dependence of Ω of radius r is shown in Figure 3.

Table 1. Values of the model parameters. Source [22]

Parameter	Value
ρ_f	1050 $[kg\ m^{-3}]$
ρ_s	1100 $[kg\ m^{-3}]$
h	42.5×10^{-6} $[m]$
E	10^{-6} $[N\ m^{-2}]$
υ	0.5 (dimensionless)
r	$[0.001, 0.01]$ $[m]$
n	2 (dimensionless)

Inserting calculated values of Ω from (8) yields the result given in Figure 4.

In order to verify that the obtained results using the model (1), the proposed model was compared with two models known in the literature:

- aneurysm as a membrane without fluid (Hung-Botwin model) [7];
- aneurysm as a mechanic aneurysm-artery system (Mast-Pierce model) [11].

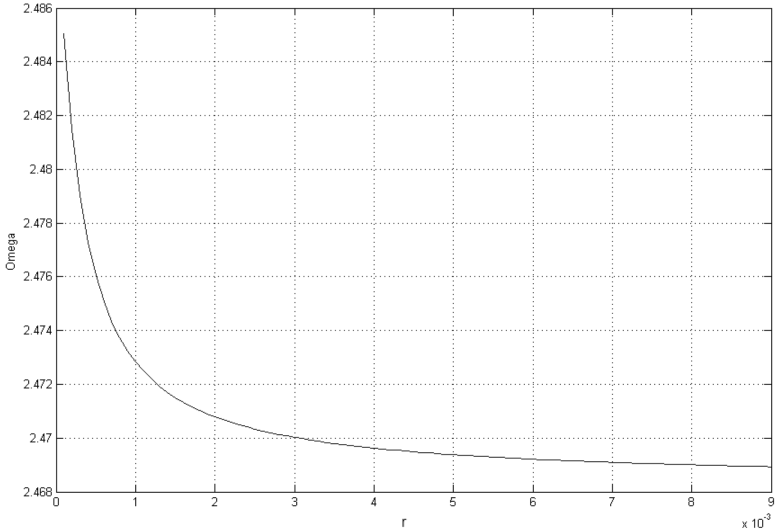


Fig. 3. Function $\Omega(r)$, for parameters in table 1

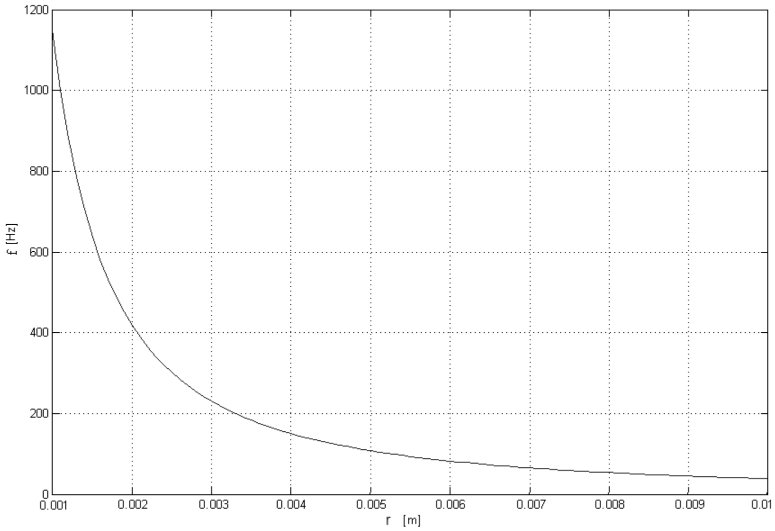


Fig. 4. Dependence f [Hz] from R [cm], for the parameters from table 1

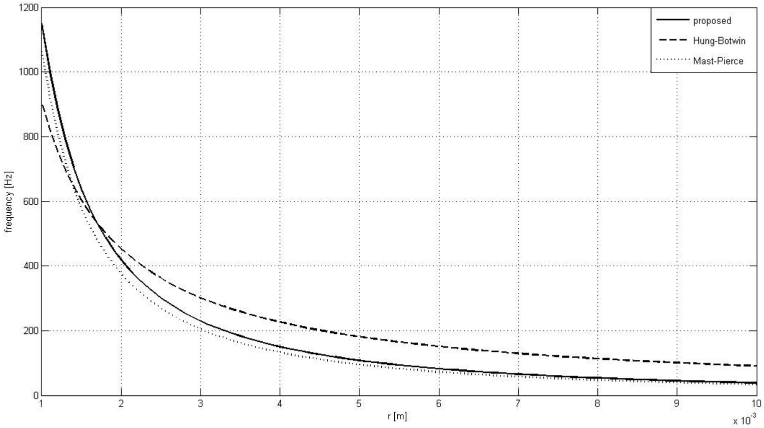


Fig. 5. Comparison of the proposed model and two models known in literature: i) aneurysm as a fluid-filled membrane [7], and ii) aneurysm as a mechanic aneurysm-artery system [11].

Comparison of the obtained results with the models known in the literature is shown in Figure 5.

3.2 Discussion

The following conclusions are drawn as a result of the frequency analysis:

1. Natural frequencies determined by the current model and the models known in literature decreases with the increase of radius.
2. The resulting natural frequency for all models were similar. However, the model of the aneurysm as a liquid-filled sphere and the Mast-Pierce model [11] (with the radius of the entrance equal to 1/2 of the radius of the aneurysm) gave almost identical results, and the Hung-Botwin model [7] (the angle of aperture 160) - very similar. These results suggest that the models are practically equivalent.
3. Comparing the audible frequencies ($100 - 1000$ Hz lub $200 - 800$ Hz) with obtained natural frequencies one can find that particularly vulnerable to the occurrence of resonance are aneurysm of size $0.01 - 0.05$ m. This result is confirmed by the statistical analysis [12], where rupture of aneurysms of that size is the most common.
4. Determining the frequency characteristics lead to the following conclusions:
   - For the smaller aneurysms the natural frequencies are higher.
   - For the smaller aneurysms the gain in resonance state is higher.

These results indicate that the resonance is possible and is equivalent to a greater risk of aneurysm rupture.

4 Experiment

This paper presents a frequency analysis of saccular intracranial aneurysms. For this purpose the model of the aneurysm as a liquid-filled spherical membrane was used. The expression for the natural frequency was derived, so one can compare the frequency determined from the model with the audible frequencies. Afterwards an experiment was set with sample values and it was compared with two models known in the literature. At the end conclusions were drawn, which clearly show that aneurysms rupture may correspond to the occurrence of the resonance.

References

1. BidzińSki, J. (ed.): Neurosurgery, Państwowy Zakład Wydawnictw Lekarskich, Warszawa (1988) (in Polish)
2. van Brugge, A.C.: The acoustic detection of intracranial aneurysms. Doctoral dissertation, Rijksuniversiteit Groningen, The Netherlands (1994)
3. Engin, A.E., King Liu, Y.: Axisymmetric response of a fluid-filled spherical shell in free vibrations. Journal of Biomechanics 3, 11–22 (1970)
4. Ferguson, G.G.: Turbulence in human intracranial saccular aneurysms. J. Neurosurg. 33, 485–497 (1970)

5. Haslach Jr., H.W.: A nonlinear dynamical mechanism for bruit generation by an intracranial saccular aneurysm. Journal of Mathematical Biology 45(5), 441–460 (2002)
6. Humphrey, J.D.: Cardiovascular Solid Mechanics. Cells, Tissues, and Organs. Springer (2002)
7. Hung, E.J.-N., Botwin, M.R.: Mechanic of rupture of cerebral saccular aneurysms. Journal of Biomechanics 8, 385–392 (1975)
8. Jain, K.K.: Mechanism of rupture of intracranial saccular aneurysms. Surg. 54, 347–350 (1963)
9. Le Roux, P.D., Winn, H.R.: Management of cerebral aneurysms – how can current management be improved. Neurosurg. Clin. of N.A. 9(3) (1998)
10. Locksley, Sahs, Knowler: In Intracranial Aneurysms and Subarachnoid Hemorrhage. Lippincott, Philadelphia (1969)
11. Mast, T.D., Pierce, A.D.: A theory of aneurysm sounds. Journal of Biomechanics 28 (1995)
12. Massachusetts Medical Society, Unruptured intracranial aneurysms – risk of rupture and risks of surgical intervention. The New England Journal of Medicine 339, 1725–1733 (1998)
13. Plett, M.I., Beach, K.W., Paun, M.: Automated ultrasonic arterial vibrometry: detection and measurement. Medical Imaging: Ultrasonic Imaging and Signal Processing (2000)
14. Sclabassi, R.J., Sun, M., Sekhar, L.N., Wasserman, J.F., Blue, H.B.: An acoustic aneurysm-detector. Medical Instrumentation 21, 317–322 (1987)
15. Sekhar, L.N., Wasserman, J.F.: Noninvasive detection of intracranial vascular lesions using and electronic stethoscope. Journal of Neurosurgery 60, 553–559 (1984)
16. Sekhar, L.N., Sun, M., Bonaddio, D., Sclabassi, R.J.: Acoustic recordings from experimental saccular aneurysms in dogs. Stroke, Journal of the American Heart Associaton 21 (1990)
17. Simkins, T.E., Stehbens, W.E.: Vibrational behavior of arterial aneurysms. Letters in Applied and Engineering Sciences 1 (1973)
18. Świątek J.:Two stage identification and its technical and biomedical applications, Wydawnictwo Politechniki Wroclawskiej, Wrocłsaw (1987) (in Polish)
19. Weir, B.: Unruptured intracranial aneurysms: a review. Journal of Neurosurgery 96 (January 2002)
20. Ziemba, S.: Vibration analysis. PWN, Warszawa (1957) (in Polish)
21. Tomczak, J.M., Gonczarek, A.: Decision rules extraction from data stream in the presence of changing context for diabetes treatment. Knowledge and Information Systems 34(3), 521–546 (2013)
22. Young, P.G.: A parametric study on the axisymmetric modes of vibration of multilayered spehrical shells with liquid cores of relevance to head impact modelling. Journal of Sound and Vibration 256(4), 665–680 (2002)

A Survey of Hardware Accelerated Methods for Intelligent Object Recognition on Camera

Aleksandra Karimaa

Business and Innovation Development,
Turku University,
Turku, Finland
aleksandra.karimaa@utu.fi

Abstract. The capability to recognize objects in online mode is an important aspect of intelligence in multimedia systems. Online object recognition provided by camera device enables video indexing to be done at camera site, which improves greatly architectural possibilities concerning material recording, search and retrieval. Classification of object at camera site enables automatic reactions concerning e.g. recording resolution or compression parameters adjustments. In multimedia systems object recognition capable camera has great potential of improving human –computer interface communication, including human-like automated decision making, i.e., automatic navigation and control tools. However, applying online object recognition requires not only efficient object recognition to be developed but they also demand near-real-time processing speed and optimization to limited resources of computational chips.

The goal of this article is to review the challenge of online object recognition on camera device, review available object recognition methods, and address their applicability in the context. Moreover, we review the issues related to using image descriptors, object definitions and object recognition in given context of online processing applied on video camera.

Keywords: Object recognition, hardware acceleration, multimedia.

1 Introduction

The amount of data generated by surrounding us systems and increasing number of possibilities of utilizing the big data, contributes to growth of research interest concerning methods of data classification, in particular when applied to content of image and video sequences. Majority of the research focuses rather on accuracy of data classification and precision of data retrieval rather than on speed of algorithms. In this work we concentrate on online object recognition focusing on speed and processing capabilities with assumption that object can be detected and recognized and classified on-fly that is without causing visible delays in video communication and with processing effort suited for embedded devices capabilities. Our goal is to identify groups of methods that will be applicable for automatic object classification in video surveillance and multimedia communication systems, i.e., camera-based object recognition.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_51, © Springer International Publishing Switzerland 2014

We review available research methods and their applicability for online object recognition and we discuss potential of hardware accelerated object recognition on camera.

2 Challenges in Object Recognition

Object recognition refers to identification and interpretation of visual perception of object. The process consists of:

- Detecting an object, e.g. there is a human in this image.
- Localizing an object, e.g. a human is located in upper left corner of the picture.
- Identifying object and its parts, e.g. human's hand belongs to human, but his walking cane is a separated object.
- Assigning object to given category, e.g. this human is an old man.
- Recognizing the object within a category, e.g. this old man is John Porter.

Whereas, for humans the process of recognition is rather simple and it greatly benefits from previously learned recognition experiences, in machine vision the process of object recognition is very complex and learning is still a challenge.

First, the process of object detection in machine vision consists of individual data processing algorithmic tasks such as object detection, tracking and recognition. These tasks are not integrated in common spatio-temporal domain. In consequence, the efficiency of machine vision, in opposite to human vision, will be greatly affected by:

- Object's occlusions and obstructions, i.e., partial visibility and disappearance of the object.
- Visual noise in the image.
- Change of view, i.e., change of object's rotation, scale, object pattern (e.g. compare the look of "houndstooth" pattern jacket from close and far distance), etc.

Second, the object recognition is never tasks in itself; it is part of challenge of understanding objects' as part of their environments. In case of human intelligence the task of interpretation of visual input is consider as rather low-complexity task (e.g. comparing to e.g. counting), but for machine intelligence such interpretation is one of most more complex system intelligence task. The object recognition task has to be integrated with tasks of intelligent interpretation of visual results in order to produce efficiency comparable with humans - which is the ultimate goal of such systems.

Third, system's ability to learn, that is to utilize knowledge from previous object recognition tasks, is still a developing area of research. Therefore, the learning aspects in the task of object recognition should not be omitted as they greatly affect the efficiency of object recognition process.

All above arguments contribute to significant computational complexity of practical implementation object recognition systems and motivate search for alternative methods to handle the complexity.

2.1 Online Object Recognition and Hardware Acceleration

The goal of this research is to survey the methods of online object recognition with potential in hardware accelerated implementation on camera device in surveillance and multimedia communication system. Most of current multimedia communication systems use High Definition (HD) or near-HD resolutions to provide the content with satisfying level of details and quality. They also use H.264 or similar encoding compression in order to fit such stream into Internet transmission channels of limited and varying capacity. We (system users) will not likely be ready to compromise good resolutions in future communication systems. Moreover, new compression algorithms are developed with aim of improving compression ratio, which imposes further growth of computational demand. A typical camera implementation today assigns entire chip processing power into one HD stream encoding. In consequence of above, we can conclude the addition of object recognition function into camera device will not be possible without hardware acceleration.

The biggest challenge related to online object detection is that the amount of computation already needed to process the data for offline object recognition is already significant and. Assuming certain level of required efficiency in object recognition we can state that redesigning object recognition algorithms with focus on improving algorithm speed will most certainly has an effect in increasing demand for computation resources. It means the process of adjusting available object recognition algorithm to be suitable for online operation will have effect on increasing computational demands by introducing more parallel processing, data caching, etc.

3 Approach

The complexity of object recognition challenge motivates us to research the object recognition by separating it into areas of potential computational challenge. This approach will allow us to find most suitable solution for surveillance and multimedia communication systems. It will also allow us to identify possible areas of future research as well as ideas of exploring less conventional methods of implementing object recognition. Moreover, we expect that good overview of methods available to address these multiple problem areas will help us to identify solutions that can be used across different areas. This, in turn, will provide us valuable insights on mentioned already spatio-temporal independency of individual algorithmic tasks. Finding such correlation different process areas might be a way to deal with challenge of limited resources when implementing object recognition on chips of camera hardware.

In principle, the problem object recognition in machine vision can be defined as a labeling problem based on models of known objects [1]. The process consists of several components:

- Searching the image to detect the objects of interests. The process is also known as image segmentation - for every image containing objects of interest and the background the process assigns labels corresponding objects to image areas. Output of this process is an object representation, which describes object properties (a

"label"). Image is then represented as a background and set of labels and their image location (specific area of the picture assigned to the label).
- Encoding of object representation to compress the information stored for each object.
- Comparing object representation (a "label") with content of object representation models. The output of this process is list of candidate object representations. Given object representation is verified against list of candidate models until the object class output is produced; such operation might require multiple iterations to be completed.

Although general process of object recognition is the same for most of object recognition implementations different approaches exist. Most variations concerns way of addressing one or more issues related to:

- Object and model representation – this represents the problem of selecting sufficient and efficient way to represent the objects. It concerns both image objects as well as model objects existing in model database. The selection of object representation will depend on type of solution, i.e., a representation of human body in medical system will require more information that representation of human body in people flow monitoring system. Information to be used to represent the object might be a subject of encoding to minimize storage, transmission and data processing times.
- Extraction of object information from the picture – in principle different object information can be searched and retrieved different way, e.g. picture might be searched for one particular color information in particular area.
- Matching retrieved information to the model object representation and model database search methods- there are different methods of comparing object data retrieved from an image with model data available in model database. That includes e.g. partial match, one-to-one, one-to-many or many-to-many [2, 3] object information matching.
- Object category verification denote different methods of environment analysis to exclude or confirm object presence in the image, e.g. an application might require an additional information from motion tracking system which will inform about probability of object being in given location. This phase is necessary to handle mentioned already problems of object occlusion in crowded environments.

3.1 Note on Feature-Based Methods

The most central decision concerning object recognition implementation is object representation. The way to represent the object has most crucial effect on final results of application efficiency and computational complexity by setting the amount of the information to be used to represent the object for further processing. The most general classification of object representation methods distinguishes between appearance and feature based methods.

Appearance-based methods represent the object in simplified way by comparing directly the image of example object with image of model object. Methods combine information about object's size, shape (edge) and color and compare it against model image to produce the decision. Methods might be suitable for high accuracy detection of small differences between similar objects, such as in medical imaging or product-line monitoring. However, the methods are not able to handle well object variances caused by e.g. different light conditions (resulting a change of color), or change of view (difference in object scale, and shape).

In contrast to appearance-based methods, the feature-based methods have advantages of being able to handle better object variances. In feature-based methods, called also model based method the object is represented by its model. Model is created by combination of object features, such as color, shape, pattern, alignment, etc. Object model allows handling correctly object variances caused by object rotation or scaling, e.g. blue hat when rotated will still remain blue. The process of object to model comparison is can be optimized by limiting number of features, limiting the areas of comparison and various feature-based conditioning, e.g. if blue color is found in this area, then search neighboring area for blue color only.

The general idea of feature-based methods is to use combination of features to describe the object and to identify the probability of object being present of the picture. It means that the methods are able to combine well other feature-like information such as not only camera-sourced features (color, shape, and pattern) with other sensor information e.g. depth [4] or distance [5] , scale from camera zoom information [5], 3D position, or background patterns information [6] and many other.

In consequence of above most of visual systems especially using movable cameras will use feature-based oriented object recognition, but combination of different features and addition of appearance-based algorithms is researched widely in search of the most optimal way to recognize the object in given applications.

4 Hardware Acceleration

Our aim is to investigate hardware acceleration aspects in context of implementation of online object recognition on camera device. We study the results of hardware implementations of these methods.

We have classified researched methods by problem areas they approach.

- Object representation methods.
- Object and object feature search methods.
- Object to model comparison and verification.

4.1 Object Representation Methods

Explored methods can be categorized into feature (model)-based methods, and hybrid methods combining both feature-based and appearance-based approaches [7].

There are various methods to describe the object. Hardware acceleration is often used in popular object representation methods:

- Object clustering methods, typically used for object with specific set of poses or shapes. Extensively used in human pose clustering. Object clustering methods provide interesting area of research providing e.g. in human pose clustering. Clustering and object segmentation is well supported by modern graphics.

- Invariance methods- object is described by set of featured invariant to given camera transformation. Geometrics information is often used for scale-invariant methods, color for rotation –invariant methods, etc. The methods might provide good results with very low implementation cost.

However, due to the fact that video surveillance need to support multi-planar camera views or multi-camera operation e.g. person tracking across different cameras' views, we have paid significant attention to hardware accelerated methods of 3D object representation.

We have identified two general approaches of object representation when it comes to hardware accelerated methods. In the first approach, model of the object is created using high quality rendering of point-sampled 3D objects. This approach has been represented e.g. by [8]. Rendering of point-sampled 3D object require hardware acceleration of rendering function. Rendering acceleration is available in modern standard PC graphics hardware, which provides good results on competitive cost. However, rendering of full 3D models might be too slow for providing 25frames per second images of reasonable resolution in speed sufficient for online object classification. In second approach, model of the object is created by calculating object models from 2D image plane and depth information calculated either from IR sensors (including Kinect devices), camera zoom information or by using stereo images [9]. In this approach the role of hardware acceleration it to speed up scene understanding by combining 2D information and depth information and create object trajectories across video sequence. Hardware support for this type of function where typical vision information is combined with information from other sources typically requires dedicated architectural solution typically it is implemented on FPGA chips [10].

4.2 Object and Object Feature Search

Search methods can be optimized by hardware acceleration concerns the one or more of following problems related to optimization of object search. The category of methods includes image descriptors methods, but also invariance and also clustering methods (described above) which by defining special object features and their descriptors provide also special support for their processing. Special research attention is given to Scale-invariant feature transform (SIFT) and Speeded Up Robust Features (SURF) methods, which are widely support in popular computing language tools.

Hardware acceleration support is also often provided often for dedicated search algorithms, such as nearest-neighbor search [11].

4.3 Object to Model Comparison and Verification

Object to model comparison is another area of computational complexity. Object to model comparison methods vary greatly depending on object representation being used. The majority of implemented methods is based on feature descriptions. In SIFT method introduced by Lowe in 2004 [12], key-points of objects are extracted from multiple reference images and object is compared to model by individually comparing image features to object's features from model database. Initial implementations of SIFT were very slow (too slow for online processing) but the speed of current implementation has been improved using parallel architectures [13, 14]. There are many examples of successful implementation of hardware embedded SIFT methods. Many of these implementations provide results sufficient for online object recognition, e.g. the implementation proposed by Huang [15] produces object recognition from one VGA image within 33ms time. SIFT method produced also improved version of method, called PCA-SIFT [16]. PCA-SIFT used features description which ware more robust to image deformation. Such method improvement allows the system to handle camera introduced object deformation. Originally proposed PCA-SIFT has also disadvantage of being slow and require additional parallelization in order to provide sufficient speed. Speed challenge of PCA-SIFT has produced SURF method [17, 18], which use the same robust features but introduces improved mechanisms of feature search and comparison. SURF implementations such as [19] or [20] proved to be successful applicant for online object recognition providing multiple object recognition from HD resolution pictures with delay not more than 40ms. Also, the methods using configurable hardware such as [21] present significant advantages of good efficiency and speed.

5 Conclusions and Future Work

Major part of current object definition research is designed in the way that the concept of object recognition uses functions available in General-purpose computing on graphics processing unit (GPGPU), which intention is to utilize the power of Graphic Processing Unit (GPU) and improve use of multiple graphic cards within one machine. Interesting alternative to developing own FPGA-based architectural solutions to satisfy application needs is presented by combination NVidia's CUDA. Sufficient performance can be provided by SIFT and SURF methods. Further improvement can be provided by further parallelization of implementations e.g. with use of OpenMP and by using configurable hardware.

Acknowledgment. The author acknowledges the contribution of Business Innovation and Development department in process of publishing this article.

References

1. Jain, R., Kasturi, R., Schunck, B.G.: Machine Vision. McGraw-Hill (1995)
2. Demirci, M.F., Shokoufandeh, A., Keselman, Y., Bretzner, L., Dickinson, S.: Object Recognition as Many-to-Many Feature Matching. Int. J. Comp. Vision 69(2), 203–222 (2006)
3. Keselman, Y., Demirci, M.F., Macrini, D., Dickinson, S.: Many-to-many feature matching in object recognition: a review of three approaches. IET Comp. Vision 6(6), 500–513 (2012)
4. Clapés, A., Reyes, M., Escalera, S.: Multi-modal User Identification and Object Recognition Surveillance System. Pattern Recognition Letters (2012)

5. Warsop, T., Singh, S.: A survey of object recognition methods for automatic asset detection in high-definition video. In: IEEE 9th Int. Conf. in Cybernetic Intelligent Systems (CIS), pp. 1–6 (2010)
6. Murphy, K., Torralba, A., Eaton, D., Freeman, W.T.: Object detection and localization using local and global features. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (eds.) Toward Category-Level Object Recognition. LNCS, vol. 4170, pp. 382–400. Springer, Heidelberg (2006)
7. Azad, P., Asfour, T., Dillmann, R.: Combining appearance-based and model-based methods for real-time object recognition and 6d localization. In: IEEE/RSJ Int. Conf. In Intelligent Robots and Systems, pp. 5339–5344 (2006)
8. Ren, L., Pfister, H., Zwicker, M.: Object space EWA surface splatting: A hardware accelerated approach to high quality point rendering. Computer Graphics Forum 21(3), 461–470 (2002)
9. Bleyer, M., Rhemann, C., Rother, C.: Extracting 3D scene-consistent object proposals and depth from stereo images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 467–481. Springer, Heidelberg (2012)
10. Becker, T., Liu, Q., Luk, W., Nebehay, G., Pflugfelder, R.: Hardware-Accelerated Object Tracking. In: Computer Vision on Low-Power Reconfigurable Architectures Workshop, Field Programmable Logic and Applications FPL (2011)
11. Bustos, B., Deussen, O., Hiller, S., Keim, D.A.: A graphics hardware accelerated algorithm for nearest neighbor search. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3994, pp. 196–199. Springer, Heidelberg (2006)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Comp. Vision 60(2), 91–110 (2004)
13. Warn, S., Emeneker, W., Cothren, J., Apon, A.: Accelerating SIFT on parallel architectures. In: IEEE Int. Conf. on in Cluster Computing and Workshops CLUSTER 2009, pp. 1–4 (2009)
14. Zhang, Q., Chen, Y., Zhang, Y., Xu, Y.: Sift implementation and optimization for multicore systems. In: IEEE Int. Symp. in Parallel and Distributed Processing IPDPS 2008, pp. 1–8 (2008)
15. Huang, F.C., Huang, S.Y., Ker, J.W., Chen, Y.C.: High-Performance SIFT Hardware Accelerator for Real-Time Image Feature Extraction. IEEE Trans. on Circuits and Systems for Video Technology 22(3), 340–351 (2012)
16. Juan, L., Gwun, O.: A comparison of sift, pca-sift and surf. Int. J.- of Image Processing (IJIP) 3(4), 143–152 (2009)
17. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
18. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3), 346–359 (2008)
19. Svab, J., Krajník, T., Faigl, J., Preucil, L.: Fpga based speeded up robust features. In: IEEE Int. Conf. on Technologies for Practical Robot Applications TePRA 2009, pp. 35–41 (2009)
20. Terriberry, T.B., French, L.M., Helmsen, J.: GPU accelerating speeded-up robust features. In: 3D Data Processing, Visualization and Transmission. Georgia Tech, Atlanta (2008)
21. Schaeferling, M., Kiefer, G.: Flex-SURF: A flexible architecture for FPGA-based robust feature extraction for optical tracking systems. In: Int. Conf. on Reconfigurable Computing and FPGAs (ReConFig), pp. 458–463 (2009)

Heuristic Solution Algorithm for Routing Flow Shop with Buffers and Ready Times

Jerzy Józefczyk and Michał Markowski

Department of Intelligent Decision Support Systems, Wrocław University of Technology,
wyb. Wyspianskiego 27, 50-370 Wrocław, Poland
{jerzy.jozefczyk,michal.markowski}@pwr.wroc.pl

Abstract. This work discusses the routing flow shop which means that jobs, located at workstations represented by nodes of a transportation network, are performed by movable machines travelling among the workstations. The version with buffers, ready times and different speeds of machines to minimize the makespan is considered. The heuristic constructive solution algorithm and its analytical assessment are presented. Results of simulation experiments evaluating the algorithm are also given.

Keywords: flow shop, routing, optimization, heuristic algorithms, computer simulation.

1 Introduction

Flow shop problems have been discussed in the literature since Johnson's article [8] in 1954 (for the recent survey see e.g. [7]). Hundreds of technical papers have been written since that time and many versions of the flow shop problem have arisen. Flow shop problems with setup times, where some setup activities have to be performed to prepare machines for next jobs, can be mentioned as the example [1], [15], [16], [13]. Another version deals with the flow shop with batches where jobs are performed in groups (see [12] for a review). The next version, which coincides with the considerations of this work, generalizes classical consideration on the case when machines can move to perform jobs. Such an assumption leads to so called routing flow shop. Investigations of this work refer also to the version with non-zero ready (release) times, see e.g. [4] (where branch and bound algorithm is proposed to solve three-machine problem with makespan as the criterion).

The flow of jobs is sometimes impossible due to the difficulties in handling or relocating jobs that are too big or too small, too heavy or cannot be moved due to technological limitations. These situations may occur for the production of ships, big wagons, cars or small parts like transistors. In these cases, moving machines can be used to drive from one job located at its workstation to another one. For example, to build a ship, four machines can work alongside the ship: the first machine polishes the surface of the ship for further processing, the second machine rivets metal plates, the third machine paints with anticorrosive paint and the last one paints the ship with

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_52, © Springer International Publishing Switzerland 2014

the final color. The parts of the ship cannot be moved due to its size; however, mobile machines can move from one area to another one. Above situation is called routing problem or problem with moving machines (executors), e.g. [2, 18].

This work is based on the considerations given by Averbakh and Berman in [2] where the simple approximation algorithm is proposed together with the evaluation of its quality for the routing flow shop problem with unlimited buffers, without ready times and with equal driving speeds of machines. Authors propose an upper bound for the solutions returned by their approximation algorithm, which will be discussed further, to solve the problem with ready times and different speeds of machines.

Jozefczyk and Markowski consider in [11] the routing flow shop problem without buffers and present constructive greedy algorithm as well as its comparison to the optimal algorithm based on a simple enumeration. Authors present also a recurrent procedure for the calculation of the makespan.

Yu et al. develop in [18] a $10/7$ approximation algorithm for two-machine routing flow shop problem, another approximation algorithm for m -machine routing open shop as well as for the routing flow shop problem with unlimited buffers, without ready times. Both algorithms deal with flow shop problem better than those presented by Averbakh et al. in [2, 3]. Yu et al. proved NP-hardiness of two-machine routing flow shop by the reduction from the well-known NP-hard partitioning problem.

Jozefczyk and Markowski present in [9] and [10] the classical task scheduling problem with routing and discuss the case with interval processing times. The objective function based on an absolute regret is used. Tabu search and simulated annealing solution algorithms are developed and compared.

Flow shop problems with routing are more complex than their classical versions because driving times and sometimes driving limitations have to be taken into account. It is important to note that flow shop with routing can be considered as the difficult and very rare investigated version of so called task scheduling with setup times and sequence dependent setups, e.g. [1].

This work is based on the considerations given in [2], where the simple approximation algorithm is proposed together with the evaluation of its quality for the routing flow shop problem with buffers, without ready times and with equal driving speeds of machines. In the paper, the generalization of the problem is investigated, which consists in taking into account ready times for jobs and different speeds of machines. A new approximate solution algorithm as well as its evaluation are presented. Moreover, the results of numerical experiments are discussed which compare the approximate algorithm proposed with the optimal algorithm being the simple enumeration.

2 Routing Flow Shop Problem

Let us consider the flow shop problem with m machines and n jobs where $\boldsymbol{M} = \{M_1, M_2, \ldots, M_r, \ldots, M_m\}$ and $\boldsymbol{J} = \{J_1, J_2, \ldots, J_h, \ldots, J_n\}$ are set of machines and set of jobs, respectively. Moreover, $r \in \{1,2,...,m\}, h \in \{1,2,...,m\}$ are indices of current machine and job. A workstation is defined as the place where a job is located. There is no particular difference between a job and a workstation, however the notion 'workstation' refers to the localization of a job. A depot as the workstation where all machines start and finish their work, and no activity is performed, is denoted by J_{n+1}. All jobs and the depot constitute set $\overline{\boldsymbol{J}} = J\{J_1, J_2, \dots, J_n, J_{n+1}\}$.

Each job is composed of m operations being parts of a job and performed by consecutive machines. Operation $O(r,h)$ refers to the part of job J_h, which is performed by machine M_r. Operations within particular job J_h are performed by machines in the given order. For the simplicity, the order of performed operations is defined hereinafter as $(O(1,h), O(2,h), \dots, O(m,h))$, $h = 1, 2, ..., n$. This order denoted as $(M_1, M_2, \dots, M_m)$ is given unlike the order of jobs undergoing the performance. Due to the movement of machines, each operation is composed of two parts: driving of a machine between workstations and performing of an operation at the workstation. We denote by $p_{r,h}$ and $\hat{p}_{r,g,h}$ the execution time of operation $O(r,h)$ and the driving time of machine M_r from workstation J_g to workstation J_h, respectively. In a consequence, $\tilde{p}_{r,g,,h} = p_{r,h} + \hat{p}_{r,g,h}$ is the execution time of operation $O(r,h)$. The ready time for job J_h denoted as u_h means that the job cannot start before this time elapses.

In order to formulate the corresponding optimization problem, the decision variable (a sequence or machines' route) is defined as follows: $\boldsymbol{S} = (J_{S_0}, J_{S_1}, J_{S_2}, \dots, J_{S_i}, \dots, J_{S_n}, J_{S_{n+1}})$, where $(S_1, S_2, ..., S_n)$ is a permutation of $(1,2,...,n)$ and $S_0 = S_{n+1} = n+1$, represents a depot. Moreover, $S_i \neq S_k$, and $S_i = h$ means that job J_h is performed as the i th.

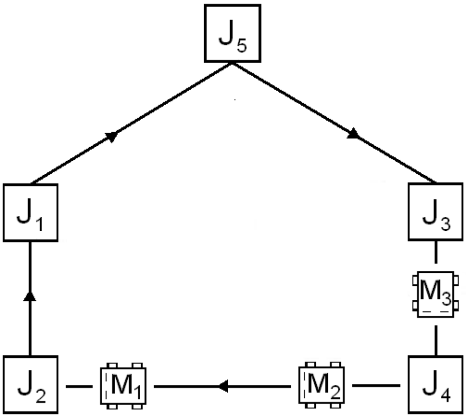


Fig. 1. Example of a layout of workstations and machine routes

This work refers to the permutation version of flow shop problem both in its classical version and with routing, e.g. [6, 14] and we assume that each machine follows the same sequence and performs jobs in the same order.

Two cases of the routing flow shop can be considered ([11]) with respect to buffers as the equipment of workstations. Workstations without buffers can only host one machine, i.e. the machine that performs a job. No additional machines are allowed to wait or stop at the workstation, where currently the operation is performed, by another machine. Such a constraint influences the calculation of the makespan. Before it drives up to the next workstation, the machine has to wait for leaving this workstation by the previous machine. This requirement does not exist in the case with buffers, which is discussed in the work. Additionally, it is assumed that buffers have unlimited capacity. The example of the routing flow shop for $n=4$, $m=3$ and $\boldsymbol{S}=(J_5, J_3, J_4, J_2, J_1, J_5)$ is presented in Fig. 1.

Makespan Calculation

Let us denote by $C(\boldsymbol{S},r,i)$ the time moment when machine M_r can start to move to the next workstation $J_{S_i} \in \overline{\boldsymbol{J}}$ where index i refers to the position in the sequence.

$$C(\boldsymbol{S},r,i) = \max[C(\boldsymbol{S},r,i-1) + \tilde{p}_{r,S_{i-2},S_{i-1}}; C(\boldsymbol{S},r-1,i) + \tilde{p}_{r-1,S_{i-1},S_i} - \hat{p}_{r,S_{i-2},S_{i-1}}] \tag{1}$$

$$C(\boldsymbol{S},r,1) = \max[C(\boldsymbol{S},r-1,1) + \tilde{p}_{r-1,S_0,S_1} - \hat{p}_{r,S_0,S_1}; 0],\ r = 2, 3, ..., m. \tag{2}$$

The start times of jobs on machine M_1 are calculated differently than for other machines. They are delayed due to the ready times u_h, i.e.

$$\begin{gathered} C(\boldsymbol{S},1,i) = \max[C(\boldsymbol{S},1,i-1) + \tilde{p}_{1,S_{i-2},S_{i-1}}; u_{S_i} - \hat{p}_{1,S_{i-1},S_i}], \\ i = 2, 3, \ldots, n+2, u_{n+1} = u_{n+2} = 0 \end{gathered} \tag{3}$$

and

$$C(\boldsymbol{S},1,1) = \max[0; u_{S_1} - \hat{p}_{1,S_0,S_1}]\,. \tag{4}$$

Finally, the makespan is the maximum of the return to the depot by all machines

$$C_{\max}(\boldsymbol{S}) = \max_{r=1,2,...,m} (C(\boldsymbol{S},r,n) + \tilde{p}_{r,S_{n-1},S_n} + \hat{p}_{r,S_n,S_{n+1}})\,. \tag{5}$$

Consequently, the routing flow shop problem considered consists in determining of such a sequence $\boldsymbol{S}$ to minimize $C_{\max}(\boldsymbol{S})$ for given: $\boldsymbol{M}, \boldsymbol{J}$, $p_{r,h}$, $\hat{p}_{r,g,h}, u_h$, $r = 1, 2, ..., m$, $g, h = 1, 2, ..., n$.

3 Solution Algorithm

Non-zero ready times and different speeds of machines distinguish the routing flop shop problem investigated from the similar one considered in [2] where the approximate solution algorithm is presented. The results given in the work are based on those reported in [2] and can be treated as their generalization. New approximation is given which follows the corresponding evaluation of the classical version of the flow shop problem.

3.1 Approximation for Classical Flow Shop Problem

Let us denote additionally

S^T – feasible solution reverse to S, i.e. $S^T = (J_{n+1}, J_n, ..., J_2, J_1, J_0)$,

$p_h = \sum_{r=1}^{m} p_{r,h}$, $p^r = \sum_{h=1}^{n} p_{r,h}$ – execution time of job J_h, time required by machine M_r for performing activities of all jobs,

$$PMAX = \max\,(\max_h p_h, \max_r p^r),\ u_{\max} = \max_h u_h\ ,$$

$\overline{C}_{\max}(\mathrm{S}), \overline{C}_{\max}^*$ – makespan (according to classic flow shop) for S where $\hat{p}_{r,g,h} = 0$, optimal makespan, respectively.

If we consider the solution algorithm - SYMM1, which gives solution S_{SYMM1}:

1. For any sequence S compute S^T and $\overline{C}_{\max}(S^T)$,
2. If $\overline{C}_{\max}(S) \le \overline{C}_{\max}(S^{\mathrm{T}})$ accept S as the solution, otherwise take S^T,

then the following lemma is true.

Lemma: For the solution obtained by algorithm SYMM1 $\overline{C}_{\max}(S_{SYMM1}) \le (\frac{m+1}{2})\overline{C}_{\max}^* + u_{\max}$.

Proof. It is obvious that $PMAX \le \overline{C}_{\max}^*$. Without a loss of generality, let us suppose that jobs are indexed according to the sequence S, i.e. $S = (J_0, J_1, ..., J_n, J_{n+1})$ which means $S_1 = 1, S_2 = 2, ..., S_n = n.$ So, there exist $j'(0), j'(1), j'(2), ..., j'(m-1), j'(m) \in \{1,2,...,n\}$ such that $1 = j'(0) \le j'(1) \le j'(2) \le ... \le j'(m-1) \le j'(m) = n$.

Then $\overline{C}_{\max}(S) = \sum_{h=1}^{j'(1)} p_{1,h} + \sum_{h=j'(1)}^{j'(2)} p_{2,h} + ... + \sum_{h=j'(m-2)}^{j'(m-1)} p_{m-1,h} + \sum_{h=j'(m-1)}^{n} p_{m,h}$

Thus, $\overline{C}_{\max}(S) = \sum_{r=1}^{m} \sum_{h=j'(r-1)}^{j'(r)} p_{r,h}$ and after introducing ready times $\overline{C}_{\max}(S) \le u_{\max} + \sum_{r=1}^{m} \sum_{h=j'(r-1)}^{j'(r)} p_{r,h}$.

Similarly, for the reverse solution $S^T = (J_{n+1}, J_n, J_{n-1}, ..., J_1, J_0)$ there exists the sequence of $j''(0), j''(1), j''(2), ..., j''(m-1), j''(m) \in \{1,2,...,n\}$ such that $n = j''(0) \ge j''(1) \ge j''(2) \ge ... \ge j''(m-1) \ge j''(m) = 1$, for which

$$\overline{C}_{\max}(S^T) = \sum_{h=j''(1)}^{n} p_{1,h} + \sum_{h=j''(2)}^{j''(1)} p_{2,h} + ... + \sum_{h=j''(m-1)}^{j''(m-2)} p_{m-1,h} + \sum_{h=1}^{j''(m-1)} p_{m,h}.$$

Hence, $\overline{C}_{\max}(S^T) = \sum_{r=1}^{m} \sum_{h=j''(r)}^{j''(r-1)} p_{r,h}$ and, after introducing ready times, $\overline{C}_{\max}(S^T) \le u_{\max} + \sum_{r=1}^{m} \sum_{h=j''(r)}^{j''(r-1)} p_{r,h}$.

When having sequence $\boldsymbol{S}$, we take the better of $\boldsymbol{S}$ and $\boldsymbol{S}^T$. Consequently,

$$\overline{C}_{\max}(\boldsymbol{S})+\overline{C}_{\max}(\boldsymbol{S}^T)\le 2u_{\max}+(m+1)PMAX \tag{6}$$

The justification of this inequality implies from the argumentation given in [2] for the case with the same ready times. From (6) we get

$$\min\,(\overline{C}_{\max}(\boldsymbol{S}),\,\overline{C}_{\max}(\boldsymbol{S}^T)=\overline{C}_{\max}(\boldsymbol{S}_{SYMM1})\le u_{\max}+(\frac{m+1}{2})PMAX\,.$$

It is also obvious that the execution time of job J_h by any machine or the execution time of all jobs by the longest working machine is always less or equal than $\overline{C}^*_{\max}$, so $\overline{C}_{\max}(\mathrm{S}_{SYMM1})\le(\frac{m+1}{2})\overline{C}^*_{\max}+u_{\max}$. ■

The following example illustrates the approximation expressed by Lemma.

Example. Let us set $m=4$, $n=6$ and execution times $\tilde{p}_{r,g,h}$ like in Table 1.

Figure 2 presents the split of the schedule by operations, highlighted in light grey, that forms the value of $\overline{C}_{\max}(\boldsymbol{S})$. Such operations $O(r,h)$ are called critical operations. Please note that if ready times are added to the schedule, the makespan is increased by not more than $u_{\max}$.

Values of both $\overline{C}_{\max}(\mathrm{S})$ and $\overline{C}^*_{\max}$ are calculated using critical operations. Such operations are marked in Table 2 by X and Y for $\boldsymbol{S}$ and $\boldsymbol{S}^T$, respectively, which structure is equivalent to Table 1. Elements of Table 2 correspond to individual jobs and machines.

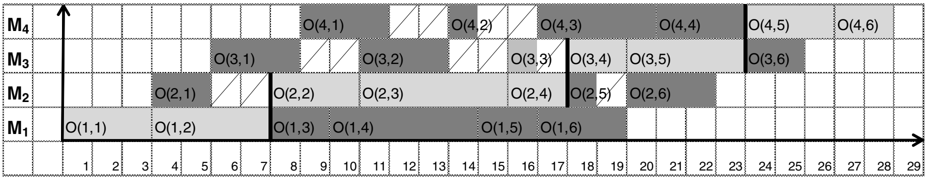
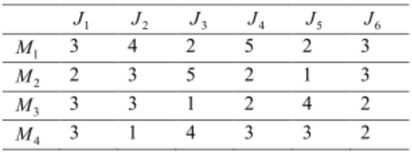


Fig. 2. Gantt chart for classical flow shop problem with buffer and zero ready times

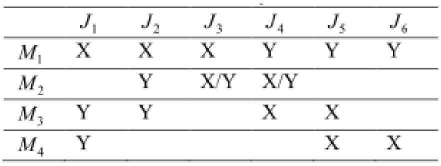
Using Table 1 and markers in Table 2, it is easy to realize that $\overline{C}_{\max}(\boldsymbol{S})+\overline{C}_{\max}(\boldsymbol{S}^{\mathrm{T}})$ is lower than the sum of corresponding times in columns plus one additional column where critical operations $O(2,3)$ and $O(2,4)$ belong both to $\boldsymbol{S}$ and $\boldsymbol{S}^T$.

Table 1. Table of execution $\tilde{p}_{r,g,h}$



	J_1	J_2	J_3	J_4	J_5	J_6
M_1	3	4	2	5	2	3
M_2	2	3	5	2	1	3
M_3	3	3	1	2	4	2
M_4	3	1	4	3	3	2

Table 2. Table with marked critical operations S and S^T



	J_1	J_2	J_3	J_4	J_5	J_6
M_1	X	X	X	Y	Y	Y
M_2		Y	X/Y	X/Y		
M_3	Y	Y		X	X	
M_4	Y				X	X

3.2 Approximation for Routing Flow Shop Problem

The routing flow shop problem can be treated as a combination of two classical subproblems: multiple traveling salesmen (MTSP) when $p_{r,h} = 0$ and flow shop when $\hat{p}_{r,g,h} = 0$. Then, let us denote as $Q(\boldsymbol{S})$ and $\overline{C}_{\max}(\boldsymbol{S})$ values of $C_{\max}(\boldsymbol{S})$ for $p_{r,h} = 0$ and $\hat{p}_{r,g,h} = 0$, respectively. Therefore,

$$C_{\max}(\boldsymbol{S}) = Q(\boldsymbol{S}) + \overline{C}_{\max}(\boldsymbol{S}) \tag{7}$$

for machines driving at the same speed. If there is a sequence of jobs in the classical flow shop problem, and if to this sequence the equal driving times, that depend only on the distances between workstations and do not depend on the machine, are added, then each job can only be delayed by the same period on each machine. This delay of jobs does not influence any other delays that can increase the makespan more than of the value $Q(\boldsymbol{S})$.

However, if machines drive at different speeds, the following expression is true

$$C_{\max}(\boldsymbol{S}) \le Q(\boldsymbol{S}) + \overline{C}_{\max}(\boldsymbol{S}) . \tag{8}$$

Inequality (8) results straightforwardly from (7). If machines drive at different speeds, value of related $Q(\boldsymbol{S})$ will be the time of visiting all workstations from $\boldsymbol{J}$ by the slowest machine. If all machines have a speed of the slowest machine then (7) will be introduced. Otherwise, if any machine drives faster than others, additional savings in time may occur in $C_{\max}(\boldsymbol{S})$ (only if the driving time will in place of the idle time that occurred), thus $C_{\max}(\boldsymbol{S}) \le Q(\boldsymbol{S}) + \overline{C}_{\max}(\boldsymbol{S})$.

If machines drive at the same speed, it is obvious that

$$C^*_{\max} \ge Q^* + \overline{C}^*_{\max} \tag{9}$$

because $Q(\boldsymbol{S}^*) \triangleq Q^*$ and $\overline{C}_{\max}(\boldsymbol{S}^*) \triangleq \overline{C}^*_{\max}$ are values of the criteria for the optimal solution $\boldsymbol{S}^*$ when respectively MTSP and flow shop are solved separately. However, if different driving speeds of machines take place, (9) may not be true. Let us consider the optimal solution of the classical flow shop that has an idle time t between two jobs J_g and J_h on machine M_r. If the driving time is added in such a way that between every two jobs the driving times equals zero except J_g and J_h on M_r, and the driving time is not more than t ($\hat{p}_{i,k,l} = 0 \forall i \ne m, k \ne g, l \ne h, p_{r,g,h} \le t$), (9) is not true. Moreover, $C^*_{\max} = \overline{C}_{\max}(\boldsymbol{S}^*)$, but $Q\ (\boldsymbol{S}^*) > 0$. To take into account such a situation, the additional time, referred to as $\hat{p}_{\max}$ should be added to $C^*_{\max}$. It is the maximum value that would increase of $C^*_{\max}$ if all machines drive at the speed of the slowest one. To calculate $\hat{p}_{\max}$ it is necessary to add the greatest of n times $\hat{p}^{g,h}_{\max}$

where $\hat{p}^{g,h}_{\max} = \max_r \hat{p}_{r,g,h} - \min_r \hat{p}_{r,g,h}$, $g,h = 1, 2, ..., n$, $g \neq h$. Finally, if machines drive at different speed

$$C^*_{\max} + \hat{p}_{\max} \geq Q^* + \overline{C}^*_{\max} . \tag{10}$$

Let S_ε be the ε –approximate solution algorithm of MTSP, i.e. $Q(S_\varepsilon) \leq (1+\varepsilon)Q^*$. Christofides's approximation algorithm of the time complexity $O(n^3)$ and $\varepsilon = \frac{3}{2}$ can serve as such an algorithm [6]. The algorithm is the basis for the following solution algorithm of the routing flow shop considered, called SYMM2.

Algorithm SYMM2:

1. Compute $C_{\max}(S_\varepsilon)$ and $C_{\max}(S_\varepsilon{}^T)$.
2. If $C_{\max}(S_\varepsilon) \leq C_{\max}(S_\varepsilon{}^T)$ $S_{\text{SYMM 2}} = S_\varepsilon$, otherwise $S_{\text{SYMM 2}} = S^{\text{T}}_\varepsilon$.

The Theorem is true where L_{C} is the lower bound of $C^*_{\max}$, for example $L_C = \max(\min_h u_h - \max_h \tilde{p}_{\breve{r},n+1,h};0) + \max_r \breve{p}_r$ where

$\breve{p}_r = (\sum_{h=1}^{n} p_{r,h} + \sum_{h=1}^{n+1} \min_g \tilde{p}_{r,g,h} \forall g = \{1,...,n+1\})$ and $\breve{r} = \arg\max_r \breve{p}_r$, $u_{\max} = \max_h u_h$.

Theorem: For the solution obtained by algorithm SYMM2

$$\frac{C_{\max}(S_{SYMM\,2})}{C^*_{\max}} \leq \max\{1+\varepsilon;\frac{m+1}{2}\}(1+\frac{\hat{p}_{\max}}{L_C}) + \frac{u_{\max}}{L_C} .$$

Proof:

The proof results immediately from (8), (10) and Lemma. Namely,

$$C_{\max}(S_{SYMM\,2}) \leq Q(S_{SYMM\,2}) + \overline{C}_{\max}(S_{SYMM\,2}) \tag{11}$$

Using Lemma, (11) can be rewritten as

$$C_{\max}(S_{SYMM\,2}) \leq Q(S_{SYMM\,2}) + \frac{m+1}{2}\overline{C}_{\max} + u_{\max} \leq (1+\varepsilon)Q^* + \frac{m+1}{2}\overline{C}^*_{\max} + u_{\max} ,$$

$$(1+\varepsilon)Q^* + \frac{m+1}{2}\overline{C}^*_{\max} + u_{\max} \leq \max\{1+\varepsilon;\frac{m+1}{2}\}(Q^* + \overline{C}^*_{\max}) + u_{\max} .$$

After introducing (10) we have

$$\max\{1+\varepsilon;\frac{m+1}{2}\}(Q^* + \overline{C}^*_{\max}) + u_{\max} \leq \max\{1+\varepsilon;\frac{m+1}{2}\}(\hat{p}_{\max} + C^*_{\max}) + u_{\max} ,$$

$$C_{\max}(S_{SYMM\,2}) \leq \max\{1+\varepsilon;\frac{m+1}{2}\}(\hat{p}_{\max} + C^*_{\max}) + u_{\max} ,$$

$$\frac{C_{\max}(S_{SYMM\,2})}{C^*_{\max}} \leq \max\{1+\varepsilon;\frac{m+1}{2}\}(\frac{\hat{p}_{\max}}{C^*_{\max}} + 1) + \frac{u_{\max}}{C^*_{\max}} .$$

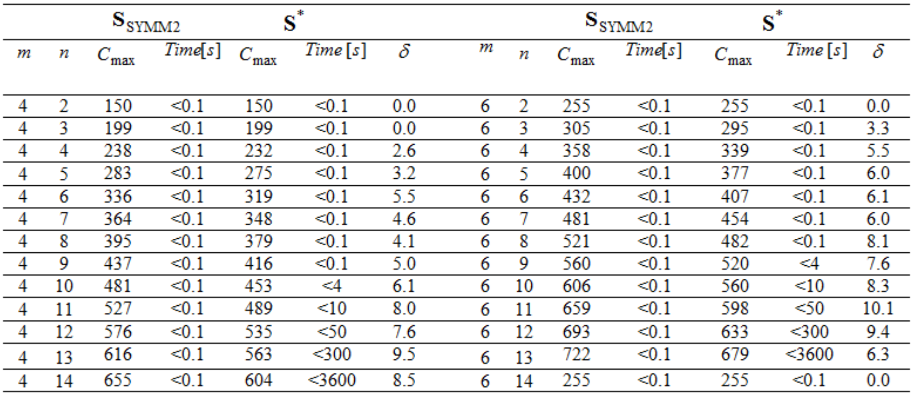
Finally:

$$\frac{C_{\max}(S_{SYMM\,2})}{C^{*}_{\max}} \le \max\{1+\varepsilon; \frac{m+1}{2}\}(\frac{\hat{p}_{\max}}{L_C}+1)+\frac{u_{\max}}{L_C}\,. \qquad \blacksquare$$

4 Simulation Experiments

The solution algorithm SYMM2 has been evaluated during simulation experiments. The performance index $\delta = \frac{C_{\max}(S_{\text{SYMM}\,2}) - C^{*}_{\max}}{C^{*}_{\max}} 100\%$ is the basis for the assessment, where $C^{*}_{\max}$ has been calculated by a simple enumeration. The further insertion ([17]) approximate algorithm has been used for S_{ε} where $\varepsilon = 2$. Values of $p_{r,h}$, $\hat{p}_{r,g,h}$ and u_h were randomly generated according to the rectangular distribution from intervals [1; 50], [1; 30], [0; 20], respectively. Results for $m \in \{4,6\}$ and $n \in \{2,3,...,13,14\}$ are given in Table 3 where computation times *Time* of the algorithms are presented. Each value of δ and *Time* is the average of 10 independent runs.

Table 3. $C_{\max}$ and δ values for the instance of routing flow shop problem



		S_{SYMM2}		S^{*}					S_{SYMM2}		S^{*}		
m	n	$C_{\max}$	*Time*[*s*]	$C_{\max}$	*Time* [*s*]	δ	m	n	$C_{\max}$	*Time*[*s*]	$C_{\max}$	*Time* [*s*]	δ
4	2	150	<0.1	150	<0.1	0.0	6	2	255	<0.1	255	<0.1	0.0
4	3	199	<0.1	199	<0.1	0.0	6	3	305	<0.1	295	<0.1	3.3
4	4	238	<0.1	232	<0.1	2.6	6	4	358	<0.1	339	<0.1	5.5
4	5	283	<0.1	275	<0.1	3.2	6	5	400	<0.1	377	<0.1	6.0
4	6	336	<0.1	319	<0.1	5.5	6	6	432	<0.1	407	<0.1	6.1
4	7	364	<0.1	348	<0.1	4.6	6	7	481	<0.1	454	<0.1	6.0
4	8	395	<0.1	379	<0.1	4.1	6	8	521	<0.1	482	<0.1	8.1
4	9	437	<0.1	416	<0.1	5.0	6	9	560	<0.1	520	<4	7.6
4	10	481	<0.1	453	<4	6.1	6	10	606	<0.1	560	<10	8.3
4	11	527	<0.1	489	<10	8.0	6	11	659	<0.1	598	<50	10.1
4	12	576	<0.1	535	<50	7.6	6	12	693	<0.1	633	<300	9.4
4	13	616	<0.1	563	<300	9.5	6	13	722	<0.1	679	<3600	6.3
4	14	655	<0.1	604	<3600	8.5	6	14	255	<0.1	255	<0.1	0.0

The inaccuracy of the approximation algorithm does not always escalate when the number of jobs n increases. After $n = 9$ inaccuracy is less than 12% and this should be analyzed in more details. Although, this inaccuracy does not exceed 17% in total, a more effective algorithm is required. Proposed upper bound value does not limit the value of $C_{\max}$ in a very tight way. There is still a room for improvement and presentation of more tight upper bound.

Further work in this area may include obtaining upper bounds for other versions of routing flow shop, i.e. without buffers and with other criteria. The absence in literature of heuristic algorithms for routing flow shop problems has been noticed.

5 Final Remarks

The generalized routing scheduling flow shop problem with buffers to minimize the makespan is considered in the paper. The generalization consists in taking into account different speeds of machines and ready times for jobs. The solution algorithm is proposed. It is based on the solution of the corresponding multiple travelling salesman problem. The analytical evaluation of the algorithm and its experimental verification using a numerical experiments are presented. It is planned to extend investigations for other versions of the routing flow shop. In particular, cases with other scheduling criteria, e.g. the maximum tardiness or the sum of completion times of jobs are worth elaborating.

References

1. Allahverdi, A., Ng, C.T., Cheng, T.C.E., Kovalyov, M.Y.: A survey of scheduling problems with setup times or costs. European Journal of Operational Research 187, 985–1032 (2008)
2. Averbakh, I., Berman, O.: A simple heuristic for m-machine flow-shop and its applications in routing-scheduling problems. Operations Research 47, 165–170 (1999)
3. Averbakh, I., Berman, O., Chernykh, I.: A 6/5 -approximation algorithm for the two-machine routing open-shop problem on a two-node network. European Journal of Operational Research 166, 3–244 (2005)
4. Cheng, J., Steiner, G., Stephenson, P.: A computational study with a new algorithm for the three-machine permutation flow-shop problem with release times. European Journal of Operational Research 130, 559–575 (2001)
5. Christofides, N.: Worst-case analysis of a new heuristic for the traveling salesman problem. Technical Report, GSIA, Carnegie-Mellon University, Pittsburgh, PA (1976)
6. Framinan, J.M., Gupta, J.N.D., Leisten, R.: A review and classification of heuristics for permutation flow-shop scheduling with makespan objective. Journal of the Operational Research Society 55, 1243–1255 (2004)
7. Gupta, J.N.D., Stafford, E.F.: Flowshop scheduling research after five decades. European Journal of Operational Research 169, 699–711 (2006)
8. Johnson, S.M.: Optimal two- and three-stage production schedules with setup times included. Naval Research Logistics Quarterly 1, 61–68 (1954)
9. Józefczyk, J., Markowski, M.: Heuristic Algorithms for Solving Uncertain Routing-Scheduling Problem. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2008. LNCS (LNAI), vol. 5097, pp. 1052–1063. Springer, Heidelberg (2008)
10. Józefczyk, J., Markowski, M.: Simulated annealing based robust algorithm for routing-scheduling problem with uncertain execution times. Information Control Problems in Manufacturing 13 (2009)
11. Józefczyk, J., Markowski, M.: Integrated Optimization Problems in Operations Research. In: Proceedings of 23rd Int. Conf. on Systems Research, Informatics and Cybernetics, Analysis and Decision Making for Complex and Uncertain Systems, Baden-Baden, Germany, vol. 1, pp. 39–43 (2011)
12. Potts, C.N., Kovalyov, M.Y.: Scheduling with batching: a review. European Journal of Operational Research 120, 228–249 (2000)

13. Rajendran, C.H., Ziegler, H.: Scheduling to minimize the sum of weighted flow time and weighted tardiness of jobs in a flowshop with sequence -dependent setup times. European Journal of Operational Research 149, 513–522 (2003)
14. Rebaine, D.: Flow shop vs. permutation shop with time delays. Computers & Industrial Engineering 48, 357–362 (2005)
15. Ruiz, R., Allahverdi, A.: No-wait flowshop with separate setup times to minimize maximum lateness. International Journal of Advanced Manufacturing Technology 35, 551–565 (2007)
16. Ruiz, R., Stutzle, T.: An iterated greedy heuristic for the sequence dependent setup times flowshop problem with makespan and weighted tardiness objectives. IRIDIA, Technical report number TR/IRIDIA/2006-02 (2006)
17. Syslo, M.M., Deo, N., Kowalik, J.S.: Algorithms of discrete optimization. In: PWN, p. 299 (1999) (in Polish)
18. Yu, W., Liu, Z., Wang, L., Fan, T.: Routing open shop and flow shop scheduling problems. European Journal of Operational Research 213, 24–36 (2011)

Common Route Planning for Carpoolers – Model and Exact Algorithm

Grzegorz Filcek and Dariusz Gąsior

Wroclaw University of Technology, Institute of Informatics, Wybrzeze Wyspianskiego 27,
50-370 Wroclaw, Poland
{grzegorz.filcek,dariusz.gasior}@pwr.wroc.pl

Abstract. The carpooling may be seen as an alternative transportation mode. It is based on common journey planning, so the costs may be reduced and so the traffic. In the consequence, it may have potentially a great positive impact on the environment (the less cars on roads the less pollution). The aim of this paper is to introduce a complete model for a common route planning for carpoolers. The problem is formulated as a multiobjective optimization task. The exact solution algorithm for finding Pareto-optimal solutions is given.

Keywords: multicriteria optimization, routing, carpooling.

1 Introduction

Carpooling is a common travelling of many people by one car. It allows not only to reduce travel costs of all participants, but also to decrease traffic [2]. However, it is necessary for travelers to have a common communication platform, which enables joint journey planning. Usually, some social services are used for such a purpose. Nevertheless, they do not support a process of automatic connecting people and finding the routes for them, which would enable possibly best association between passengers and drivers. Consequently, some dedicated applications are developed. However, their capabilities are still limited, because there is still lack of the algorithms, which enable taking into account all fundamental real life carpooling problem requirements [5].

The task of common route planning for carpoolers may be seen as a multicriteria multipath task. Similar problems were considered in [1, 3, 4, 6, 7, 9, 10, 15, 16, 17, 18, 20]. However, despite results presented in [11], the algorithms given in the mentioned papers concern finding one path (i.e. the problem is solved only from one person's point of view). Moreover, neither constraints concerning the times of travel beginning and completion nor the dependencies between paths of travellers (drivers and passengers) are considered.

In this paper, the complex model for the common route planning for carpoolers is presented. The exact solution algorithm is also proposed and some implementation aspects are considered. It is also indicated that due to problem complexity, the exact

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_53, © Springer International Publishing Switzerland 2014

algorithm is not time efficient. But, it is important to develop such an algorithm as a reference for further solutions.

2 Mathematical Model

In this paper it is considered that a map of the area that carpoolers operates may be given as a graph $G^{\mathrm{KP}} =< V^{\mathrm{KP}}, E^{\mathrm{KP}} >$ (where V^{KP} defines a set of nodes and E^{KP} reflects set of links). It is assumed that there are *K* drivers who travel from the origin node $\underline{v}_k^{\mathrm{K}}$ to their destination $\overline{v}_k^{\mathrm{K}}$. They are willing to reduce their travel cost by offering a carpool service. They may take some additional passengers if only their routes belong to subgraph $G_k^{\mathrm{K}} =< V_k^{\mathrm{K}}, E_k^{\mathrm{K}} >\subseteq G^{\mathrm{KP}}$ and some additional requirements concerning the capacity of their cars $(S_k \in \mathbf{N}_+)$, the longest acceptable travel time (T_k^{CK}), longest acceptable distance (L_k^{K}) and biggest acceptable travel cost $(C_k^{\mathrm{K}} \in \mathbf{R}_+)$ are satisfied. There are also *P* passengers who eager to go with drivers offering a carpool service. Any *p*th passenger may start his trip only in one of nodes belonging to $\underline{V}_p^{\mathrm{P}} \subseteq V^{\mathrm{KP}}$ and must finish his trip in a node belonging to $\overline{V}_p^{\mathrm{P}} \subseteq V^{\mathrm{KP}}$. We assume without loss of generality that every set of nodes associated with every passenger is disjunctive with any other set of nodes associated with any other passenger. Furthermore, it is supposed that any drivers' starting or final vertex does not belong to any set of nodes (either starting or finals) associated with any passenger. Besides, passengers define their requests concerning the longest acceptable travel time (T_p^{CP}), the longest acceptable travel distance between the origin node $\underline{v}_p$ and the destination node $\overline{v}_p$ (i.e., $L_p^{\mathrm{P}}(\underline{v}_p, \overline{v}_p)$) and the biggest acceptable travel cost $(C_p^{\mathrm{P}} \in \mathbf{R}_+)$.

For the sake of simplicity, we introduce the following incidence matrices: $\overline{e}_{k,i1,i2}^{\mathrm{K}} = 1(0)$ if there exist an arc in G_k^{K} connecting i_1th node with i_2th node (otherwise) and $\overline{e}_{i1,i2}^{\mathrm{KP}} = 1(0)$ if there exists an arc in G^{KP} connecting i_1th node with i_2th node (otherwise).

Moreover, there are weights related to every node and every arc in G^{KP}, reflecting properties like travelling time, distance, cost, etc. So, we introduce the following variables:

$\mathbf{W} = [\mathbf{w}_r]_{r=1,2,\ldots,\left|E^{\mathrm{KP}}\right|}$ - the vector of vectors of attributes for every arc in G^{KP},

$\mathbf{w}_r = [w_{r,kp,m}]_{\substack{kp=1,2,\ldots,K+P \\ m=1,2,\ldots,M}}$ - the vector of M attributes for *r*th arc in G^{KP},

$\overline{\mathbf{W}} = [\overline{\mathbf{w}}_d]_{d=1,2,\ldots,|V^{\mathrm{KP}}|}$ - the vector of vectors of attributes for every node in G^{KP},

$\overline{\mathbf{w}}_d = [\overline{w}_{d,kp,\overline{m}}]_{\substack{kp=1,2,\ldots,\underline{K}+P \\ \overline{m}=1,2,\ldots,\overline{M}}}$ -the vector of $\overline{M}$ attributes for *d*th node in G^{KP}.

Since the drivers and the passengers have some prerequisites concerning the moments of beginning and finishing trip, so the following variables concerning each node in graph G^{KP} are introduced:

$\underline{T}(v_d^{\mathrm{KP}})$ – the earliest departure time from the node v_d^{KP}, i.e. $\overline{w}_{d,1,1} = \overline{w}_{d,2,1} = \ldots = \overline{w}_{d,K+P,1} = \underline{T}(v_d^{\mathrm{KP}})$,

$\overline{T}(v_d^{\mathrm{KP}})$ – the latest arrival time to the node v_d^{KP} $(\overline{w}_{d,1,2} = \overline{w}_{d,2,2} = \ldots = \overline{w}_{d,K+P,2} = \overline{T}(v_d^{\mathrm{KP}}))$.

In this paper, we focus on the three aspects of travelling: time, distance and cost, so for each edge in graph G^{KP} the following weights are considered:

$T(e_n^{\mathrm{KP}})$ – travel time through edge e_n^{KP} $(w_{r,1,1} = w_{r,2,1} = \ldots = w_{r,K+P,1} = T(e_n^{\mathrm{KP}}))$

$L(e_n^{\mathrm{KP}})$ – length of edge e_n^{KP} (travel distance between nodes connected by edge e_n^{KP}) $(w_{n,1,2} = w_{n,2,2} = \ldots = w_{n,K+P,2} = L(e_n^{\mathrm{KP}}))$

$C(e_n^{\mathrm{KP}})$ – cost of traveling through edge e_n^{KP} $(w_{n,1,3} = w_{n,2,3} = \ldots = w_{n,K+P,3} = L(e_n^{\mathrm{KP}}))$

Further in this work, we use $\{1,2,3,\ldots,\mathrm{Z}\} \overset{\Delta}{=} \overline{1,\mathrm{Z}}$ as a notation for a set of natural numbers from 1 to Z.

The first decision to make to satisfy as good as possible all the drivers and passengers aims is the assignment of passengers to the drivers by indicating the nodes that form the set of activity nodes. Activity nodes set (**AC**) includes nodes of graph G^{KP}, in which the driver begins or ends the route or have to stop because of picking up or dropping off a passenger. Let us denote a binary decision variable $y_{k,d} \in \{0,1\}$ describing such assignment of the *d*th node of graph G^{KP} to the set of activity nodes (**AC**) in the path of the *k*th driver, element of the matrix $\mathbf{y} = [y_{k,d}]_{\substack{k=1,2,\ldots,K \\ d=1,2,\ldots,|V^{\mathrm{KP}}|}}$. Let us introduce the following constraints for decision matrix $\mathbf{y}$

$$\forall_{k\in\overline{1,K}}(\forall_{p\in\overline{1,P}}(\textstyle\sum_{d:v_d^{\mathrm{KP}}\in\underline{V}_p\cap V_k^{\mathrm{K}}} y_{k,d} = \sum_{d:v_d^{\mathrm{KP}}\in\overline{V}_p\cap V_k^{\mathrm{K}}} y_{k,d} \leq 1) \wedge \forall_{d:v_d^{\mathrm{KP}}\in\{\underline{v}_k^{\mathrm{K}}\}\cup\{\overline{v}_k^{\mathrm{K}}\}}(y_{k,d} = 1)) \tag{1}$$

$$\forall_{k\in\overline{1,K}}\Big(\sum\nolimits_{d:v_d^{\mathrm{KP}}\in V_k^{\mathrm{K}}\setminus(\{\underline{v}_k^{\mathrm{K}}\}\cup\{\overline{v}_k^{\mathrm{K}}\})}(y_{k,d})=0\Big), \tag{2}$$

$$\forall_{p\in\overline{1,P}}\Big(\sum_{k=1}^{K}\sum\nolimits_{d:v_d^{\mathrm{KP}}\in \underline{V_p}\cap V_k^{\mathrm{K}}}(y_{k,d})\le 1\Big), \tag{3}$$

where (1) assures, that for every passenger who travels with the *k*th driver, exactly one of his origin nodes and exactly one of his destination nodes is included in the *k*th driver's path as well as the origin and destination nodes of this driver. Constraint (2) excludes from **AC** of the *k*th driver's path all driver's nodes which are not his origin nor destination. The assignment of each passenger to no more than exactly one driver is reached when (3) is satisfied.

Before a travel path for each driver can be obtained, there is another decision to be made, which is the appropriate order of nodes in *k*th driver's **AC** to visit. This order builds the overriding path for the driver, which cannot have any cycle. To model this decision let us introduce a binary decision variable $z_{k,d1,d2}\in\{0,1\}$ describing the existence of path from node $v_{d1}^{\mathrm{K}}\in V^{\mathrm{KP}}$ to $v_{d2}^{\mathrm{K}}\in V^{\mathrm{KP}}$ which belong to the *k*th driver's **AC**, element of matrix $\mathbf{z}=[z_{k,d1,d2}]_{\substack{k=1,2,\ldots,K\\ d1=1,2,\ldots,|V^{\mathrm{KP}}|\\ d2=1,2,\ldots,|V^{\mathrm{KP}}|}}$. The decision matrix $\mathbf{z}$ must in this case satisfy the following constraints:

$$\forall_{k\in\overline{1,K}}\Big(\forall_{d1\in\overline{1,|V^{\mathrm{KP}}|}}\Big(\sum_{d2=1}^{|V^{\mathrm{KP}}|}(y_{k,d1}y_{k,d2}(z_{k,d1,d2}-z_{k,d2,d1}))\Big)\Big)=\begin{cases}1 & \text{for} \quad v_{d1}^{\mathrm{KP}}=\underline{v}_k^{\mathrm{K}}\\ -1 & \text{for} \quad v_{d1}^{\mathrm{KP}}=\overline{v}_k^{\mathrm{K}}\\ 0 & \quad \text{otherwise}\end{cases}), \tag{4}$$

$$\forall_{k\in\overline{1,K}}\Big(\forall_{d1\in\overline{1,|V^{\mathrm{KP}}|}}\Big(\sum_{d2=1}^{|V^{\mathrm{KP}}|}(1-y_{k,d1})z_{k,d1,d2}=\sum_{d2=1}^{|V^{\mathrm{KP}}|}(1-y_{k,d1})z_{k,d2,d1}=0\Big)\Big), \tag{5}$$

$$\forall_{k\in\overline{1,K}}\Big(\forall_{d1\in\overline{1,|V^{\mathrm{KP}}|}\setminus\{d:v_d^{\mathrm{KP}}=\underline{v}_k^{\mathrm{K}}\}}\Big(\sum_{d2=1}^{|V^{\mathrm{KP}}|}y_{k,d1}z_{k,d2,d1}=1\Big)\Big), \tag{6}$$

where (4) describes the law of preservation of flow from each driver's origin to it's final destination. Constraint (5) assures, that there are no routes from or to the nodes that do not belong to the **AC** of any driver. To assure, that $\mathbf{z}$ describes only a simple path between nodes in each driver's **AC**, (6) is introduced.

Finally, the final path for each driver can be obtained. Let us denote by $x_{k,d1,d2,d3,d4}\in\{0,1\}$ a binary decision variable describing the existence of edge $(v_{d3}^{\mathrm{KP}},v_{d4}^{\mathrm{KP}})$ in the *k*th driver's path between nodes v_{d1}^{KP} and v_{d2}^{KP} belonging to *k*th driver's **AC**, element of matrix $\mathbf{x}=[x_{k,d1,d2,d3,d4}]_{\substack{k=1,2,\dots,K\\ d1=1,2,\dots,|V^{\mathrm{KP}}|\\ d2=1,2,\dots,|V^{\mathrm{KP}}|\\ d3=1,2,\dots,|V^{\mathrm{KP}}|\\ d4=1,2,\dots,|V^{\mathrm{KP}}|}}$. Let us formulate appropriate constrains for matrix $\mathbf{x}$:

$$\forall_{k\in\overline{1,K}}(\forall_{d1,d2,d3\in\overline{1,|V^{\mathrm{KP}}|}}(\sum_{d4=1}^{|V^{\mathrm{KP}}|}(x_{k,d1,d2,d3,d4}-x_{k,d1,d2,d4,d3})= \\ =\begin{cases}1 & \text{for} \quad d3=d1\wedge z_{k,d1,d2}=1\\ -1 & \text{for} \quad d3=d2\wedge z_{k,d1,d2}=1))\\ 0 & \qquad\text{otherwise}\end{cases} \tag{7}$$

$$\forall_{k\in\overline{1,K}}(\forall_{d1,d2\in\overline{1,|V^{\mathrm{KP}}|}}(\forall_{d3\in\overline{1,|V^{\mathrm{KP}}|}:d3\neq d2}(\sum_{d4=1}^{|V^{\mathrm{KP}}|}(x_{k,d1,d2,d3,d4})=\begin{cases}1 & \text{for} \quad z_{k,d1,d2}=1\\ 0 & \text{for} \quad z_{k,d1,d2}=0\end{cases}))), \tag{8}$$

$$\forall_{k\in\overline{1,K}}(\forall_{d1,d2,d3,d4\in\overline{1,|V^{\mathrm{KP}}|}}(x_{k,d1,d2,d3,d4}\leq \overline{e}^{\mathrm{KP}}_{d3,d4})), \tag{9}$$

where (7) assures the preservation of flow between nodes v^{KP}_{d1} and v^{KP}_{d2}. Constraint (8) assures, that each driver's path is combined only from a simple subpaths (no cycles) connecting consecutive nodes in the overriding path described by $\mathbf{z}$. The dependence between introduced variables is as follows: $y_{k,d}=\max_{d2}\{z_{k,d,d2};z_{k,d2,d}\}=\max_{d2,d3,d4}\{x_{k,d,d2,d3,d4};x_{k,d2,d,d3,d4}\}$, $z_{k,d1,d2}=\max_{d3,d4}\{x_{k,d1,d2,d3,d4}\}$.

Let us introduce an index $\alpha_{k,p,d}\in\{0,1\}$, that describes if the *p*th passenger is in the *k*th driver's car before reaching by this car node v^{KP}_{d}. The index $\alpha_{k,p,d}$ for each driver *k* is evaluated as follows:

$$\forall_{k\in\overline{1,K}}(\forall_{p\in\overline{1,P}}(v^{\mathrm{KP}}_{d}=\underline{v}^{\mathrm{K}}_{k}\Rightarrow\alpha_{k,p,d}=0)), \tag{10}$$

$$\forall_{k\in\overline{1,K}}(\alpha_{k,p,d}=\sum_{d1=1}^{|V^{\mathrm{KP}}|}((\alpha_{k,p,d1}+\sum\nolimits_{d2:v^{\mathrm{KP}}_{d2}\in\underline{V}_{p}}(I_{|V^{\mathrm{KP}}|})_{d1,d2}-\sum_{d2:v^{\mathrm{KP}}_{d2}\in\overline{V}_{p}}(I_{|V^{\mathrm{KP}}|})_{d1,d2})\cdot z_{k,d1,d})), \tag{11}$$

$$\forall_{k\in\overline{1,K}}(\forall_{p\in\overline{1,P}}(\forall_{i\in\overline{1,|V^{\mathrm{KP}}|}}(\alpha_{k,p,d}\geq 0))). \tag{12}$$

To assure that the number of passengers who travel with one driver will never exceed the capacity of the driver's car the following constraint is given

$$\forall_{k\in\overline{1,K}}(\forall_{d\in\overline{1,\left|V^{\mathrm{KP}}\right|}}(\sum_{p=1}^{P}\alpha_{k,p,d}\leq S_k)). \tag{13}$$

The decision must also satisfy passengers and drivers constraints concerning travel time, distance and costs:

$$\forall_{k\in\overline{1,K}}(\forall_{d1,d2:z_{k,d1,d2}=1}(\forall_{d3,d4:x_{k,d1,d2,d3,d4}=1}(\overline{T}(v_{\mathrm{d4}}^{\mathrm{KP}})-\underline{T}(v_{\mathrm{d3}}^{\mathrm{KP}})\geq T((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}}))))) \tag{14}$$

$$\forall_{k\in\overline{1,K}}(\sum\nolimits_{d1,d2,d3,d4:x_{k,d1,d2,d3,d4}=1}\max\{\underline{T}(v_{\mathrm{d3}}^{\mathrm{KP}})+T((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}})),\underline{T}(v_{\mathrm{d4}}^{\mathrm{KP}})\}-\underline{T}(v_{\mathrm{d3}}^{\mathrm{KP}})\leq T_{\mathrm{k}}^{\mathrm{CK}}) \tag{15}$$

$$\forall_{p\in\overline{1,P}}(\forall_{k\in\overline{1,K}}(\sum\nolimits_{(d1,d2,d3,d4:x_{k,d1,d2,d3,d4}=1\wedge\alpha_{k,p,d2}=1)}\max\{\underline{T}(v_{\mathrm{d3}}^{\mathrm{KP}})+T((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}})),\underline{T}(v_{\mathrm{d4}}^{\mathrm{KP}})\}-\underline{T}(v_{\mathrm{d3}}^{\mathrm{KP}})\leq T_{\mathrm{p}}^{\mathrm{CP}})) \tag{16}$$

$$\forall_{p\in\overline{1,P}}(\forall_{k\in\overline{1,K}}(\sum\nolimits_{(d1,d2,d3,d4:x_{k,d1,d2,d3,d4}=1\wedge\alpha_{k,p,d2}=1)}L((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}}))\leq L_p^{\mathrm{P}})) \tag{17}$$

$$\forall_{k\in\overline{1,K}}(\sum\nolimits_{(d1,d2,d3,d4:x_{k,d1,d2,d3,d4}=1)}L((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}}))\leq L_k^{\mathrm{K}})) \tag{18}$$

$$\forall_{p\in\overline{1,P}}(\forall_{k\in\overline{1,K}}(\sum\nolimits_{(d1,d2,d3,d4:x_{k,d1,d2,d3,d4}=1\wedge\alpha_{k,p,d2}=1)}C((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}}))/(1+\sum_{p=1}^{P}\alpha_{k,p,d2})\leq C_p^{\mathrm{P}})) \tag{19}$$

$$\forall_{k\in\overline{1,K}}(\sum_{(d1,d2,d3,d4:x_{k,d1,d2,d3,d4}=1)}C((v_{\mathrm{d3}}^{\mathrm{KP}},v_{\mathrm{d4}}^{\mathrm{KP}}))/(1+\sum_{p=1}^{P}\alpha_{k,p,d2})\leq C_k^{\mathrm{K}})) \tag{20}$$

The decision variables are evaluated by using cost function of each passenger denoted by $Q_p^{\mathrm{P}}(\mathbf{x})$, cost function of each driver denoted by $Q_k^{\mathrm{K}}(\mathbf{x})$, and utility function of the environment denoted as $Q^{\mathrm{S}}(\mathbf{x})$. The last function describes how much the realization of a decision can influence the environment. The cost functions of carpoolers are linear combinations of the sums of attributes (e.g. travelling time, distance, cost) along the path. In this paper we propose to thread cost functions of the passengers and drivers as equivalent, and utility function of the environment as discriminator of decision, which are non-dominated (Pareto optimal) for *P*+*K* criteria vector of passengers and drivers utility functions. So, the decision-making problem can be formulated as the following multicriteria optimization problem:
Problem **P1**:

For the given data: $G^{\mathrm{KP}},\mathbf{W},\overline{\mathbf{W}},G_k^{\mathrm{K}},T_k^{\mathrm{CK}},L_k^{\mathrm{K}},C_k^{\mathrm{K}},\underline{V}_p^{\mathrm{P}},\overline{V}_p^{\mathrm{P}},T_p^{\mathrm{CP}},L_p^{P},C_p^{P}$, determine non-dominated $\mathbf{x}$ feasible with respect to constraints (1)-(20) to minimize cost functions of all passengers and drivers, i.e.

$$\min_{\mathbf{x}}\{Q_1^{\mathrm{P}}(\mathbf{x}),Q_2^{\mathrm{P}}(\mathbf{x}),...,Q_P^{\mathrm{P}}(\mathbf{x}),Q_1^{\mathrm{K}}(\mathbf{x}),Q_2^{\mathrm{K}}(\mathbf{x}),...,Q_K^{\mathrm{K}}(\mathbf{x})\} \tag{21}$$

Problem **P2**:

For the given data: $G^{\mathrm{KP}}, \mathbf{W}, \overline{\mathbf{W}}, G_k^{\mathrm{K}}, T_k^{\mathrm{CK}}, L_k^{\mathrm{K}}, C_k^{\mathrm{K}}, \underline{V}_p^{\mathrm{P}}, \overline{V}_p^{\mathrm{P}}, T_p^{\mathrm{CP}}, L_p^{P}, C_p^{P}$, determine $\mathbf{x}$ feasible with respect to constraints (1)-(20) to maximize the utility function of environment for non-dominated decisions being solutions of the multicriteria optimization problem **P1**, i.e.

$$\max_{\mathbf{x}} Q^{\mathrm{S}}(\arg\min_{\mathbf{x}}\{Q_1^{\mathrm{P}}(\mathbf{x}), Q_2^{\mathrm{P}}(\mathbf{x}), \ldots, Q_P^{\mathrm{P}}(\mathbf{x}), Q_1^{\mathrm{K}}(\mathbf{x}), Q_2^{\mathrm{K}}(\mathbf{x}), \ldots, Q_K^{\mathrm{K}}(\mathbf{x})\}) \quad (22)$$

3 Solution Algorithm

The introduced model imposes that the problem may be decomposed into few subproblems. First of all one must find assignment between drivers and passengers. Then for each passenger it must be chosen which node is his origin and which is the destination. Then each driver orders the origin and destination nodes of all passengers assigned to him. Finally, the Pareto-optimal paths between each consecutives nodes are found and they are composed in the set of Pareto-optimal paths for each driver. However, all this problems cannot be solved separately, since there are dependencies between them. It is proposed to generate all feasible solution, which may belong to the Pareto set. Finally, the solution of the problem **P1** may be summarized as follows:

1. Generate A – a set of all possible assignments of passengers to drivers (it is possible that a passenger is not assigned to any driver, then his cost function is constant).
2. For each assignment a in A, generate B_a – a set containing all possible sequences of origin-destination nodes pair for all passengers.
3. For each a in A and for each b in B_a and for each driver k, find F_{abk} – set of all feasible sequences of nodes which must be traversed by particular driver (which defines the order of passing the origin or destination nodes in b).
4. For each a in A and for each b in B and for each driver k and for each f in F_{abk} find the R_{abkfi} – a set of all pareto-optimal subpaths between two consecutive nodes in f.
5. For each a in A and for each b in B and for each driver k and for each f in F_{abk} find R_{abkf} – a set of all possible paths from the driver's origin node to destination composed from the subpaths from R_{abkci}.
6. For each a in A and for each b in B and for each driver k and for each f in F_{abk} find R^*_{abkf} – a subset of R_{abkf} containting only non-dominated paths.
7. Choose only such elements a in A, correlated elements b in B_a, f in F_{abk} for all k and r in R_{abkf} such that set of paths (r) for all drivers (k) is non-dominated by any other set of paths for all drivers for different a, b, f.

Obviously, this algorithm may be seen as a template. In practice, each step must be executed using appropriate algorithm.

In Step 1 one obtain possible values of $y_{k,d}$ which satisfies condition (3). Then, all possible pairs of $\underline{v}_p$ and $\overline{v}_p$ are found. In the third step, the feasible values of $\mathbf{z}$ fulfilling constraints (4)-(6) are determined. It must be stressed that it is crucial to find effective way of determining feasible sequence of nodes in Step 3. Since one must remember that the origin node of each passenger must precede his destination node and the capacity of the driver's car is finite.

In Step 4, one must find solution $\mathbf{x}$. In this step, the algorithm for multiobjective shortest path problem must be implemented. Some effective algorithms based on evolutionary approach [8, 19] and simulated annealing [12] are already elaborated. Nevertheless, they are insufficient for the finding exact solution, so in this paper, application of label correction approach [7] is proposed. It is worth noting that once it is calculated a Pareto-optimal solution between two nodes for any traveler, it may be used each time, there has to be found a Pareto-optimal solution for another carpooler. That is because if the solution is non-dominated for a set of objectives, it is also non-dominated for a set of objectives defined as linear combinations with positive coefficients of the primary objectives. So, in this case, it is enough to consider only sums of individual attributes along the path as primary objectives.

It can be proven that a Pareto-optimal path must be composed only from Pareto-optimal subpaths. However, not every path composed from Pareto-optimal subpaths must be Pareto-optimal. That is why non-dominated paths have to be chosen in Step 6.

4 Final Remarks

In this paper, the model of the common route planning for the carpoolers is introduced. The problem is formulated as the multiobjective optimization task, which may be seen as an extension of multicriteria shortest multipath problem. The exact algorithm is given. However, due to time complexity, it is rather not suitable for practical application, but may be used in future works to evaluate other algorithms. For future works it is planned to implement NSGA-II algorithm, which usually gives promising results for such complex problems.

References

1. Qian, Z., Michael Zhang, H.: Modeling multi-modal morning commute in a one-to-one corridor network. Transportation Research Part C: Emerging Technologies 19(2), 254–269 (2011)
2. Abrahamse, W., Keall, M.: Effectiveness of a web-based intervention to encourage carpooling to work: A case study of Wellington, New Zealand. Transport Policy 21(0), 45–51 (2012)
3. Androutsopoulos, K.N., Zografos, K.G.: Solving the multi-criteria time-dependent routing and scheduling problem in a multimodal fixed scheduled network. European Journal of Operational Research 192(1) (2009)

4. Bozkurt, S., Yazici, A., Keskin, K.: A multicriteria route planning approach considering driver preferences. Presented at the 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp. 324–328 (2012)
5. Cho, S., Yasar, A.-U.-H., Knapen, L., Bellemans, T., Janssens, D., Wets, G.: A Conceptual Design of an Agent-based Interaction Model for the Carpooling Application. Procedia Computer Science 10, 801–807 (2012)
6. Granat, J., Guerriero, F.: The interactive analysis of the multicriteria shortest path problem by the reference point method. European Journal of Operational Research 151(1), 103–118 (2003)
7. Guerriero, F., Musmanno, R.: Label correcting methods to solve multicriteria shortest path problems. Journal of Optimization Theory and Applications 111(3), 589–613 (2001)
8. Herbawi, W., Weber, M.: Evolutionary multiobjective route planning in dynamic multi-hop ridesharing. In: Hao, J.-K. (ed.) EvoCOP 2011. LNCS, vol. 6622, pp. 84–95. Springer, Heidelberg (2011)
9. Horn, M.E.T.: Multi-modal and demand-responsive passenger transport systems: a modelling framework with embedded control systems. Transportation Research Part A: Policy and Practice 36(2), 167–188 (2002)
10. Jozefowicz, N., Semet, F., Talbi, E.-G.: Multi-objective vehicle routing problems. European Journal of Operational Research 189(2), 293–309 (2008)
11. Knapen, L., Keren, D., Yasar, A.-U.-H., Cho, S., Bellemans, T., Janssens, D., Wets, G.: Analysis of the Co-routing Problem in Agent-based Carpooling Simulation. Procedia Computer Science 10, 821–826 (2012)
12. Liu, L., Mu, H., Luo, H., Li, X.: A simulated annealing for multi-criteria network path problems. Computers & Operations Research 39(12), 3119–3135 (2012)
13. Liu, L., Mu, H., Yang, X., He, R., Li, Y.: An oriented spanning tree based genetic algorithm for multi-criteria shortest path problems. Applied Soft Computing 12(1), 506–515 (2012), doi:10.1016/j.asoc.2011.08.015
14. Mishra, S., Welch, T.F., Jha, M.K.: Performance indicators for public transit connectivity in multi-modal transportation networks. Transportation Research Part A: Policy and Practice 46(7), 1066–1085 (2012)
15. Nadi, S., Delavar, M.R.: Multi-criteria, personalized route planning using quantifier-guided ordered weighted averaging operators. International Journal of Applied Earth Observation and Geoinformation 13(3), 322–335 (2011)
16. Niaraki, A.S., Kim, K.: Ontology based personalized route planning system using a multi-criteria decision making approach. Expert Systems with Applications 36(2, pt. 1), 2250–2259 (2009)
17. Opasanon, S., Miller-Hooks, E.: Multicriteria adaptive paths in stochastic, time-varying networks. European Journal of Operational Research 173(1), 72–91 (2006)
18. Pajor, T.: Multi-modal route planning. Master thesis, Univ. Karlsruhe, TH (2009)
19. Pangilinan, J.M.A., Janssens, G.K.: Evolutionary Algorithms for the Multiobjective Shortest Path Problem. International Journal of Computer and Information Engineering 1(5), 322–327 (2007)
20. Yu, H., Lu, F.: A Multi-modal Route Planning Approach With An Improved Genetic Algorithm. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 38, Part II (2010)

Allocation-Pricing Game for Multipath Routing in Virtual Networks

Magdalena Turowska, Dariusz Gąsior, and Maciej Drwal

Institute of Computer Science,
Wrocław University of Technology,
Wrocław, Poland
{magdalena.turowska,dariusz.gasior,maciej.drwal}@pwr.wroc.pl

Abstract. In this paper we consider the problem of routing data packets in virtual network. It is assumed that each virtual link may consist of arbitrary collection of paths composed of physical links. Traffic originating from a source node may reach its destination using many paths simultaneously. In order to maximize the utility in a network where a large number of such transmissions occur, it is necessary to allocate bandwidth of physical links appropriately among all component flows, as well as to decide on the division of capacity of virtual links. We apply game-theoretic analysis to the formulated multipath routing problem in order to develop a distributed mechanism for solving network utility maximization problem in virtual network. We show that the obtained allocations form a strategy profile which is a pure Nash equilibrium. Moreover, the proposed mechanism allows to allocate the bandwidth in a way which maximizes the total utility.

Keywords: computer networks, game theory, mechanism design.

1 Introduction

In the last years virtualization technologies are becoming increasingly more popular among network operators, which need to deploy various data access services efficiently without expensive hardware upgrades [28], [6]. The use of software-emulated networking devices allows managing the access to the shared network resources in a flexible device-independent manner. While virtualization of computer systems is currently well-understood and widespread, network virtualization is still an active research direction. The idea is to logically separate the network links' capacity, in order to set up independently operating sub-networks, designated to serve a particular higher-level application. Such virtual networks can be formed into a particular topologies, designed to fulfill the needs of specific groups of users. The creation and reconfiguration of such network can be performed automatically and resources which are no longer needed can be released for the purpose of future demands.

In the seminal work of [19] the use of utility theory interpretation of the flow control algorithms in large scale packet-switched communication networks was

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_54, © Springer International Publishing Switzerland 2014

introduced. This approach allowed to model mathematically the transmission control protocol (TCP), which is a fundamental building block of the Internet [19], [17], [31]. Recently, it became a standard tool for designing and analyzing other rate control algorithms in terms of utility functions. These ideas have been widely adopted in many works, e.g. in [18], [2], [21], [22], [10], [29], [32], [23], [14], [12], [8] and [13]. A recent summary of the state of the art is provided in [3].

The adaptation of the utility theory approach to the analysis of control laws and design of algorithms for the rate allocation in virtual networks appeared for the first time in [15]. In [7] we provided the rate control and capacity allocation algorithms for the special case of network with one level of virtualization. An extensive overview and current research directions in the area of network virtualization may be found i.a. [1], [33], [6], [4], [11] and [5].

Motivated by the decentralized nature of the Internet, and pervasive selfishness of interacting agents, the research in the area of algorithmic game theory [27] have become very prolific in the last years. The game theory is a powerful framework for studying decision making problems involving a group of agents acting individually, being rational and competing or cooperating to achieve certain goals [24]. Game-theoretic analysis of rate allocation problem can be found in [30]. Application of auction-based mechanism for this problem can be found in [16]. Some introductory material on game theory may be found in [9], [26] while we refer to [25] as a comprehensive textbook on algorithmic game theory.

In this paper we investigate the problem of routing packet flows in computer network, where each flow's transmission rate is determined independently. The origin-destination pair is considered as decision-making agent, who wishes to maximize its utility selfishly. We assume that a single flow (that is, a transmission between two nodes in network) may consist of multiple component flows. Such multi-component flow can be seen as a single virtual network, serving a particular application, and can be set up on demand. In each such network there is a single destination node (e.g. application server), and a collection of sources (e.g. group of users). The goal is to allocate the limited links' capacity for all component flows, in order to maximize the sum of utilities of all decision-making agents. While in general allocations may include zero transmission rates, this can be seen as a problem of multipath routing.

We provide a game-theoretic model of such routing problem for independent virtual networks. In order to prevent the degradation of total utility due to the selfishness, we design a mechanism, which achieves a pure Nash equilibrium of rate allocations. Subsequently, we show that this equilibrium is in fact the socially optimal allocation of network capacity.

2 Problem Formulation

In this section we provide a formulation of the *network utility maximization* (NUM) problem in virtual network. In this problem, physical network links' capacity is divided into a collection of virtual links, and such links act as a medium for carrying users' traffic (see Figure 1). Network operators decide on

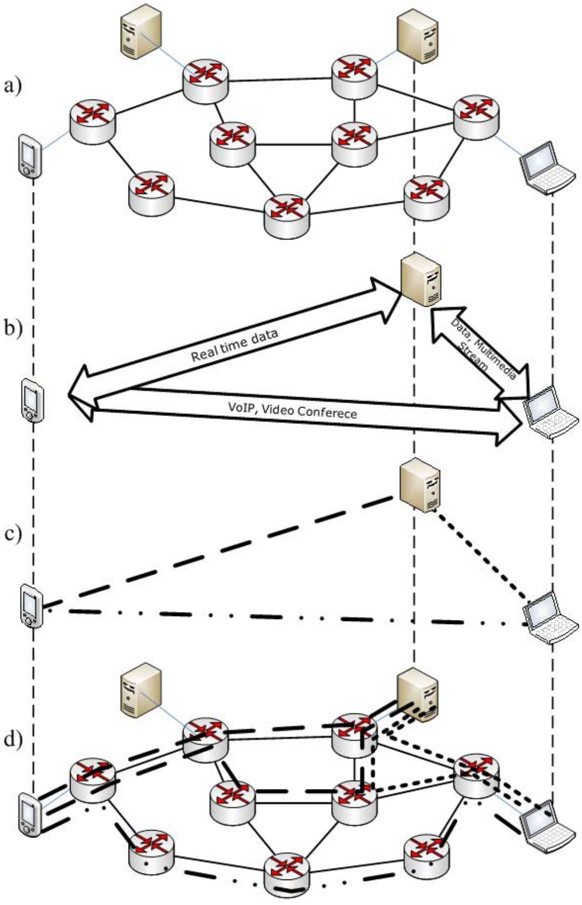


Fig. 1. Concept of network virtualization. Virtual links are considered logical connections between pairs of nodes. Within each virtual link a collection of paths carry the actual packet traffic.

the allocation of capacity among virtual links, which constrains the allocation of rates of user flows. The goal is to select such user transmission rates so as to achieve the highest total utility. In order to present the formal statement of the problem, we need to define our network model.

2.1 Flow Network Model

We consider a communication network composed of L *physical links* and N nodes, with a topology modeled by a directed graph given by a matrix $A = [a_{nl}] \in \mathbb{R}^{N \times L}$, whose entry $a_{nl} = 1$ if $l \in O(n)$, $a_{nl} = -1$ if $l \in I(n)$ and $a_{nl} = 0$ otherwise, where $I(n)$ and $O(n)$ are the collections of incoming and outgoing physical links from node n, respectively.

Between each pair of source node n and destination node d it is possible to create a *virtual link*. Within such link there are K_{nd} user flows, characterized by a transmission rate $x_{ndk} \geq 0$, $k = 1, \ldots, K_{nd}$. The utility of user flow is measured with the use utility function $F_{ndk}(x_{ndk})$, which reflects the willingness-to-pay of user for provided transmission services. We assume that F_{ndk} belongs the the class of iso-elastic utility functions [3], given by:

$$F_{ndk}(x) = \begin{cases} w_{ndk} \frac{1}{1-\gamma} x^{1-\gamma} & \gamma > 0, \gamma \neq 1, \\ w_{ndk} \log x & \gamma = 1, \end{cases} \tag{1}$$

where w_{ndk} and γ are nonnegative parameters.

The aggregate transmission rate of all user flows is denoted by:

$$x_{nd} = \sum_{k=1}^{K_{nd}} x_{ndk}.$$

Value x_{nd} determines the capacity of virtual link.

Let us define the utility of virtual link as the maximal profit that link's operator can achieve from allocating aggregate rate x_{nd}:

$$F_{nd}(x_{nd}) = \max_{\sum_k x_{ndk} = x_{nd}} \sum_{k=1}^{K_{nd}} F_{ndk}(x_{ndk}), \tag{2}$$

consequently, from (1) we have for $\gamma > 0, \gamma \neq 1$:

$$F_{nd}(x_{nd}) = \left(\sum_{k=1}^{K_{nd}} w_{ndk}^{1/\gamma} \right)^{\gamma} (1-\gamma) x_{nd}^{(1-\gamma)},$$

and for $\gamma = 1$:

$$F_{nd}(x_{nd}) = \left(\sum_{k=1}^{K_{nd}} w_{ndk} \right) \log x_{nd} + \sum_{k=1}^{K_{nd}} \left[w_{ndk} \log \left(w_{ndk} / \sum_{q=1}^{K_{nd}} w_{ndq} \right) \right].$$

On each physical link l, $y_{dl} \geq 0$ is the amount of all aggregated flows destined to node d. At each node n, these quantities satisfy the flow preservation constraint:

$$\forall_{d=1,\ldots,N} \, \forall_{n=1,\ldots,N} \quad x_{nd} + \sum_{l \in I(n)} y_{dl} = \sum_{l \in O(n)} y_{dl}.$$

We assume limited capacity of links:

$$\forall_{l=1,\ldots,L} \quad \sum_{d=1}^{N} y_{dl} \leq C_l,$$

where C_l is the capacity of link l.

The problem of determining capacity allocations for virtual links can be interpreted as the problem of multipath routing. It is assumed that each aggregated flow in virtual link can be transmitted to destination node d simultaneously through a collection of paths originating from different nodes n.

2.2 Network Utility Maximization

Many network resource allocation problems can be formulated as constrained maximization of some utility function. In the *network utility maximization* (NUM) approach [20], each transmission demand (or user) has its utility function and link bandwidths are allocated so that network utility is maximized. Efficiency of global resource allocation can be measured by the total network utility (i.e. the sum of virtual link utilities), representing so-called *social welfare*.

The problem of optimal design of virtual network, stated as multipath routing problem in physical links, can be formulated as the following continuous optimization problem in variables x_{nd} and y_{dl}:

$$\text{maximize} \quad \sum_{d=1}^{N} \sum_{n=1, n\neq d}^{N} F_{nd}(x_{nd}) - \sum_{l=1}^{L} \sum_{d=1}^{N} P_l(y_{dl}) \tag{3}$$

subject to:

$$\forall_{d=1,\ldots,N} \, \forall_{n=1,\ldots,N} \quad x_{nd} + \sum_{l \in I(n)} y_{dl} = \sum_{l \in O(n)} y_{dl}, \tag{4}$$

$$\forall_{d=1,\ldots,N} \, \forall_{n=1,\ldots,N} \, \forall_{l=1,\ldots,L} \quad x_{nd} \geq 0, y_{dl} \geq 0, \tag{5}$$

$$\forall_{l=1,\ldots,L} \quad \sum_{d=1}^{N} y_{dl} \leq C_l. \tag{6}$$

Function F_{nd} represents the utility of transmission from source n to destination d, and function P_l represents the cost of usage of link l to transmit data, as measured by the rate of allocated flow.

Observe that given a solution $(\mathbf{x}, \mathbf{y})$ of this problem, which amounts to the allocation of physical capacity to virtual links, it is easy to determine the optimal rates of user flows x_{ndk} within the virtual links, by solving the maximization in the expression (2) for F_{nd}. For the assumed class of iso-elastic utilities (1) this amounts to evaluating the expression for each $k = 1, \ldots, K_{nd}$:

$$x_{ndk} = x_{nd} \frac{w_{ndk}^{1/\gamma}}{\sum_{q=1}^{K_{nd}} w_{ndk}^{1\gamma}}.$$

3 Problem Decomposition

Since F_{nd} are a concave functions, and we assume that P_l are strictly convex and differentiable functions, the objective function (3) is strictly concave and differentiable. The feasible region is compact. Hence, values of x_{nd} and y_{dl} maximizing the objective function exist and can be found by Lagrange method. The Lagrangian of formulated problem is:

$$L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{d=1}^{N} \sum_{n=1, n\neq d}^{N} F_{nd}(x_{nd}) + \sum_{l=1}^{L} \lambda_l \left(\sum_{d=1}^{N} y_{dl} - C_l \right) +$$

$$+\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}\beta_{nd}\left(\sum_{l\in O(n)}y_{dl}-\sum_{l\in I(n)}y_{dl}-x_{nd}\right)-\sum_{l=1}^{L}\sum_{d=1}^{N}P_l(y_{dl})=$$

$$=\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}F_{nd}(x_{nd})-\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}\beta_{nd}x_{nd}+\sum_{d=1}^{N}\sum_{l=1}^{L}\lambda_l y_{dl}+$$

$$+\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}\sum_{l=1}^{L}\beta_{nd}a_{nl}y_{dl}-\sum_{l=1}^{L}\lambda_l C_l-\sum_{l=1}^{L}\sum_{d=1}^{N}P_l(y_{dl}),$$

where λ_l, β_{nd} are Lagrange multipliers (prices) associated with link capacities and network structure constraints, respectively.

This convex optimization problem (optimization a concave function over a convex constraint set) has useful Lagrange duality properties, which lead to decomposable structures. Lagrange duality theory links the original maximization problem, termed primal problem, with a dual minimization problem, which sometimes readily presents decomposition possibilities. The basic idea in Lagrange duality is to relax the original problem by transferring the constraints to the objective in the form of a weighted sum.

We use dual decomposition methods for resource allocation. The master problem sets the price for the resources to each subproblem, which has to decide the amount of resources that can be used depending on the price.

The Lagrange dual problem associated with the primal problem under consideration is given by:

$$\text{minimize}\quad L(\mathbf{x},\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\beta}) \tag{7}$$

in variables $\boldsymbol{\lambda},\boldsymbol{\beta}$, subject to a single constraint: $\forall_{l=1,\dots,L}\ \ \lambda_l\geq 0$, where:

$$L(\mathbf{x},\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\beta})=L_x(\mathbf{x},\boldsymbol{\beta})+L_y(\mathbf{y},\boldsymbol{\lambda})$$

and:

$$L_x(\mathbf{x},\boldsymbol{\beta})=\max_{\mathbf{x}}\left(\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}F_{nd}(x_{nd})-\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}\beta_{nd}x_{nd}\right),$$

$$L_y(\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\beta})=\max_{\mathbf{y}}\left[\sum_{l=1}^{L}\lambda_l\left(\sum_{d=1}^{N}y_{dl}-C_l\right)+\sum_{d=1}^{N}\sum_{n=1,n\neq d}^{N}\sum_{l=1}^{L}\beta_{nd}a_{nl}y_{dl}\right.$$

$$\left.-\sum_{l=1}^{L}\sum_{d=1}^{N}P_l(y_{dl})\right].$$

It is well known that for a convex optimization, a local optimum is also a global optimum and solving the problem after decomposition is equivalent to solving the original (primal) problem. Consequently, solving the original problem reduces to solving two types of problems:

A) the determination of x_{nd} by n-th transmission source:

$$\text{maximize} \quad [F_{nd}(x_{nd}) - x_{nd}\beta_{nd}] ,$$

B) the determination of y_{dl} by the l-th physical link:

$$\text{minimize} \quad \left[P_l(y_{dl}) - \lambda_l \left(\sum_{d=1}^{N} y_{dl} - C_l \right) - \sum_{n=1}^{N} \beta_{nd} a_{nl} y_{dl} \right] .$$

The Lagrange multipliers can be determined by solving (7) for given fixed $\mathbf{x}$ and $\mathbf{y}$. This dual problem can be solved in a distributed way with a subgradient method, resulting in updated Lagrange multipliers.

As results of the decomposition we have obtained two subproblems: subproblem A, which can be solved by users (sources) independently, and subproblem B, which can be solved independently by physical links.

4 Game-Based Mechanism

The problem of computing optimal solution of network utility maximization problem by distributed independent agents suffers from the effect of selfishness. As each decision agent is interested only in maximizing their own utility, we can expect that their decisions would eventually stabilize at a solution that cannot be unilaterally improved. However, such solution can be far from the optimal allocation in terms of social welfare (i.e. objective of NUM).

In order to prevent selfish behaviors of users we introduce a coordination mechanism for subproblem A. We formulate a game-theoretic model of interaction of users. We refer to the class of noncooperative games derived from such mechanism as auction, in which a Nash equilibrium is usually assumed as a basic solution concept. The Nash equilibrium captures the notion of a stable solution and is the solution from which no decision agent (player) can individually deviate in order to improve utility.

In this paper we assume that each player is restricted to communicate his demand to the network with the price that the player is willing to pay for the resources. We present a distributed and dynamic process that converges to Nash equilibrium.

Let us consider the following network game: each virtual link (n, d) is associated with one player who determines the aggregate transmission rate x_{nd} and the price β_{nd} as:

$$x^*_{nd} \in \arg\max_{x_{nd}} \left[F_{nd}(x_{nd}) - x_{nd}\bar{\beta}_{nd} \right] , \tag{8}$$

and:

$$\beta^*_{nd} = \bar{\beta}_{nd} \left[1 - \eta(z_{nd} - x^*_{nd}) \right] . \tag{9}$$

Both formulas require the computation of quantity $\bar{\beta}_{nd}$, which equals the mean value of all players' prices β_{nd}, except that of player (n, d):

$$\bar{\beta}_{nd} = \frac{1}{M-1} \left(-\beta_{nd} + \sum_{j=1}^{N} \sum_{k=1, k\neq j}^{N} \beta_{jk} \right) ,$$

where $M = \sum_{j=1}^{N} \sum_{i=1, i \neq j}^{N} K_{ji}$ is the number of players, $\eta > 0$ is a parameter of algorithm and:

$$z_{nd} = \sum_{l \in O(n)} y_{dl} - \sum_{l \in I(n)} y_{dl} = \sum_{l=1}^{L} a_{nl} y_{dl}.$$

This strategy is a consequence of application of the following service cost:

$$q_{nd}(x_{nd}, \beta_{nd}) = x_{nd}\bar{\beta}_{nd} + \left[\beta_{nd} - \bar{\beta}_{nd}(1 - \eta(z_{nd} - x_{nd}))\right]^2,$$

which is the price that player pays for rate x_{nd} corrected by the penalty that player pays due to the mismatch between its price and the average price of the other players.

Each player determines a strategy (x_{nd}, β_{nd}) by solving:

$$(x_{nd}^*, \beta_{nd}^*) \in \arg \max_{x_{nd}, \beta_{nd}} \left[F_{nd}(x_{nd}) - q_{nd}(x_{nd}, \beta_{nd})\right] =$$

$$= \arg \max_{x_{nd}, \beta_{nd}} \left[F_{nd}(x_{nd}) - x_{nd}\bar{\beta}_{nd} - (\beta_{nd} - \bar{\beta}_{nd}(1 - \eta(z_{nd} - x_{nd})))^2\right]. \quad (10)$$

The maximization of this net utility can only occur when (x_{nd}^*, β_{nd}^*) are computed as in (8) and (9), respectively. The function $F_{nd}(x_{nd}) - x_{nd}\bar{\beta}_{nd}$ is the payoff function of player associated with data transmission from source n to destination d.

The determination of y_{dl} and λ_l can be performed in the same way as described in Subsection 3.

5 Properties

In this section we show that the mechanism presented in previous section implements the solution of the problem described in Subsection 2.2. It can be shown that Nash equilibrium of mechanism presented in Section 4 satisfy the optimality conditions for original problem (3)–(6). The necessary and sufficient conditions for the efficient allocation can be determined from Karush-Kuhn-Tucker (KKT) conditions.

Remark 1. At a Nash equilibrium, the value of (10) is maximized for each player.

Theorem 1. *The game mechanism* (8)–(9) *implements the problem* (3)–(6) *in Nash equilibrium.*

Proof. The first order optimality conditions for (10) are:

$$\frac{\partial F_{nd}(x_{nd})}{\partial x_{nd}} - \bar{\beta}_{nd} - \frac{\partial q_{nd}(x_{nd}, \beta_{nd})}{\partial x_{nd}} = 0 \quad (11)$$

and

$$\frac{\partial q_{nd}(x_{nd}, \beta_{nd})}{\partial \beta_{nd}} = 0. \quad (12)$$

From (12) we obtain:

$$\beta_{nd} - \bar{\beta}_{nd}\left[1 - \eta\left(z_{nd} - x_{nd}\right)\right] = 0.$$

After substitution to (11):

$$\frac{\partial F_{nd}(x_{nd})}{\partial x_{nd}} = \bar{\beta}_{nd},$$

which gives:

$$\beta_{nd} = \frac{1}{M-1}\left(-\beta_{nd} + \sum_{j=1}^{N}\sum_{k=1,k\neq j}^{N}\beta_{jk}\right)\left[1 - \eta(z_{nd} - x_{nd})\right].$$

This implies that:

$$\forall_{n,d} \quad z_{nd} - x_{nd} = 0, \tag{13}$$

and

$$\forall_{n,d} \quad \beta_{nd} = \bar{\beta}_{nd} = \beta.$$

Consequently, we obtain:

$$\forall_{n,d} \quad \frac{\partial F_{nd}(x_{nd})}{\partial x_{nd}} = \bar{\beta}_{nd} = \beta,$$

and from (13):

$$\forall_{n,d} \quad \sum_{l\in I(n)} y_{dl} + x_{nd} = \sum_{l\in O(n)} y_{dl}.$$

Similarly for the part associated with determination of y_{dl} we receive:

$$\frac{\partial(P_l(y_{dl}) - \lambda_l y_{dl} - \sum_{n=1}^{N}\beta_{nd}a_{nl}y_{dl})}{\partial y_{dl}} = \frac{\partial P(y_{dl})}{\partial y_{dl}} - \lambda_l - \sum_{n=1}^{N}\beta_{nd}a_{nl} = 0,$$

for $\beta_{nd} = \beta$:

$$\lambda_l = \frac{\partial P(y_{dl})}{\partial y_{dl}}$$

and

$$\frac{\partial L(\boldsymbol{\lambda},\boldsymbol{\beta})}{\partial \lambda_l} = \sum_{d=1}^{N} y_{dl} - C_l = 0.$$

The obtained equations constitute the first order KKT conditions of the problem (3)–(6) from Section 2.2.

□

6 Conclusions

The presented mechanism allows to compute optimal solution of the network utility maximization problem in virtual network, where users' utilities depend on the allocated rates of collections of simultaneous flows on physical level. The presented procedure is based on local computations, however, it requires the exchange of price parameters β_{nd} between each pair agents in order to reach an equilibrium. The technique is suitable for virtual network management, where the performance depends critically on the appropriate allocation of virtual link capacity.

Acknowledgment. This research is co-financed by the European Union as part of the European Social Fund.

References

1. Anderson, T., Peterson, L., Shenker, S., Turner, J.: Overcoming the Internet impasse through virtualization. Computer 38(4), 34–41 (2005)
2. Chiang, M.: Nonconvex Optimization for Communication Networks. In: Advances in Applied Mathematics and Global Optimization, pp. 1–60 (2009)
3. Chiang, M., Low, S.H., Calderbank, A.R., Doyle, J.C.: Layering as optimization decomposition: A mathematical theory of network architectures. Proceedings of the IEEE 95(1), 255–312 (2007)
4. Chowdhury, N., Boutaba, R.: Network virtualization: state of the art and research challenges. IEEE Communications Magazine 47(7), 20–26 (2009)
5. Chowdhury, N., Boutaba, R.: A survey of network virtualization. Computer Networks 54(5), 862–876 (2010)
6. Creeger, M.: Moving to the edge: a CTO roundtable on network virtualization. Communications of the ACM 53(8), 55–62 (2010)
7. Drwal, M., Gasior, D.: Utility-based rate control and capacity allocation in virtual networks. In: Proceedings of the 1st European Teletraffic Seminar, pp. 176–181 (2011)
8. Drwal, M., Jozefczyk, J.: Decomposition algorithms for data placement problem based on lagrangian relaxation and randomized rounding. Annals of Operations Research (2013)
9. Easley, D., Kleinberg, J.: Networks, crowds, and markets. Cambridge University Press (2010)
10. Gasior, D.: QoS Rate Allocation in the Computer Network Under Uncertainty. Kybernetes 37(5), 693–712 (2008)
11. Gasior, D.: Capacity allocation in multilevel virtual networks under uncertainty. In: Proc. of the XV International Telecommunications Network Strategy and Planning Symposium (2012)
12. Gąsior, D., Drwal, M.: Decentralized algorithm for joint data placement and rate allocation in content-aware networks. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2012. CCIS, vol. 291, pp. 32–44. Springer, Heidelberg (2012)
13. Gasior, D., Drwal, M.: Pareto-optimal Nash equilibrium in capacity allocation game for self-managed networks. Arxiv preprint arXiv:1206.2448 (2012)

14. Gasior, D., Drwal, M.: Two-level heuristic algorithm for utility-based data placement and rate allocation in content-aware networks. In: IEEE/IPSJ 12th International Symposium on Applications and the Internet (SAINT), pp. 308–313 (2012)
15. He, J., Zhang-Shen, R., Li, Y., Lee, C.Y., Rexford, J., Chiang, M.: Davinci: Dynamically adaptive virtual networks for a customized internet. In: Proc. of the 2008 ACM CoNEXT Conference, pp. 1–12. ACM (2008)
16. Jain, R., Varaiya, P.: Efficient market mechanisms for network resource allocation. In: 44th IEEE Conference on Decision and Control, 2005 and 2005 European Control Conference, CDC-ECC 2005, pp. 1056–1061. IEEE (2005)
17. Kelly, F.: Fairness and stability of end-to-end congestion control. European Journal of Control 9(2-3), 159–176 (2003)
18. Kelly, F., Voice, T.: Stability of end-to-end algorithms for joint routing and rate control. ACM SIGCOMM Computer Communication Review 35(2), 5–12 (2005)
19. Kelly, F.P.: Mathematical modelling of the Internet. In: Mathematics Unlimited-2001 and Beyond, pp. 685–702 (2001)
20. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society 49(3), 237–252 (1998)
21. Lee, J.W., Chiang, M., Calderbank, R.A.: Jointly optimal congestion and contention control based on network utility maximization. IEEE Communications Letters 10(3), 216–218 (2006)
22. Low, S.H.: A duality model of TCP and queue management algorithms. IEEE/ACM Transactions on Networking 11(4), 525–536 (2003)
23. Turowska, M.: Application of network utility maximization to joint routing and rate control in computer networks. In: Information Systems Architecture and Technology: New Developments in Web-Age Information Systems (2010)
24. Myerson, R.B.: Game theory: analysis of conflict. Harvard University Press (1997)
25. Nisan, N.: Algorithmic game theory. Cambridge University Press (2007)
26. Osborne, M.J.: An introduction to game theory. Oxford University Press (2004)
27. Papadimitriou, C.H.: Algorithms, games, and the internet. In: Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, pp. 749–753 (2001)
28. Rixner, S.: Network virtualization: Breaking the performance barrier. Queue 6(1), 36–ff (2008)
29. Tang, A., Wang, J., Low, S.H., Chiang, M.: Equilibrium of heterogeneous congestion control: Existence and uniqueness. IEEE/ACM Transactions on Networking 15(4), 824–837 (2007)
30. Yaïche, H., Mazumdar, R.R., Rosenberg, C.: A game theoretic framework for bandwidth allocation and pricing in broadband networks. IEEE/ACM Transactions on Networking (TON) 8(5), 667–678 (2000)
31. Ye, L., Wang, Z., Che, H., Chan, H.B.C., Lagoa, C.M.: Utility function of TCP. Computer Communications 32(5), 800–805 (2009)
32. Zhang, J., Zheng, D., Chiang, M.: The impact of stochastic noisy feedback on distributed network utility maximization. IEEE Transactions on Information Theory 54(2), 645–665 (2008)
33. Zhu, Y., Ammar, M.: Algorithms for assigning substrate network resources to virtual network components. In: Proc. IEEE INFOCOM, vol. 2 (2006)

Smart-Metering-Oriented Routing Protocol over Power Line Channel

Ahmed A. Elawamry, Ayman M. Hassan, and Salwa Elramly

Benha University, Orange International Centers,
Ain Shams University, Cairo
ahmedelawamry@bhit.bu.edu.eg, ayman.hassan@orange.com,
sramlye@netscape.net
http://www.feng.bu.edu.eg/

Abstract. This paper proposes an ad hoc routing algorithm to increase link reliability in power line communications over low-tension power grid. The algorithm assumes data concentrator (DC) located at the distribution transformer, which is polling meters connected to the power line and send information about energy consumption, loading profile and any other crucial data to the utility. The proposed algorithm is designed to keep the required processing complexity at the meter side to the minimum, while shifting the intelligence towards the DC. The protocol accounts for asymmetric characteristics of the power line channel, where some nodes could suffer very bad downlink quality due to noise at the meter side. These nodes couldn't receive data sent from DC and/or other nodes and are therefore classified as deaf nodes, although their transmission could be received properly by adjacent nodes. Furthermore, special packet structure is proposed to minimize algorithm overhead and packet routing mechanism. The protocol performance is compared against LOADng, LOADng-CTP and AODV in terms of protocol overhead, end-to-end delay, packet delivery ratio and memory requirements.

Keywords: PLC, AODV, LOADng, LOADng-CTP, DC, smart metering, routing protocol.

1 Introduction

Electrical power lines have been used as a communication medium extensively over the past two decades. Power Line Communication is an attractive alternative to utility companies as it provides low-investment medium for smart grid services including: smart metering, load survey, load shedding and profiling. Typical configuration comprises data concentrator (DC) located at the distribution transformer, which communicates with meters at households via low-tension distribution grid. However, as it has never been designed with communication aspects in mind, the power line channel introduces tough challenges to the communication system designer in order to achieve reliable link with acceptable availability, throughput and reachability. Factors like attenuation, narrow-band

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_55, © Springer International Publishing Switzerland 2014

and impulsive noise, and impedance variability are among the issues that affects the link quality dramatically and therefore its impact should be mitigated. At the physical layer level, coding, interleaving and noise cancellation are commonly used to enhance the channel quality. At the network layer level, ad hoc routing protocols are used to achieve the same goal. In this arrangement, intermediate meters act like repeaters to regenerate packets from/to meters that couldn't be reached directly by the DC due to bad channel conditions. The routing algorithm should be designed to avoid network flooding at large number of nodes (meters), using a simple algorithm with small memory requirements to fit easily within the meter circuitry.

In this paper, a Low Complexity ad hoc routing protocol that is optimized for Smart Metering application (LCSM) is introduced, simulated, evaluated and compared to similar routing protocols, specifically: AODV, LOADng and LOADng-CTP protocols. LCSM protocol is designed to keep the required routing rules at the meter as simple as possible, while shifting the processing and memory requirements to the DC, where cost increase could be much more tolerated. Different Key Performance Indicators (KPIs) are evaluated including routing overhead, end-to-end delay, packet delivery ratio, topology discovery time and memory requirements. OPNET network simulator is used to evaluate the performance of LCSM protocol against AODV, LOADng and LOADng-CTP routing protocols.

The rest of the paper is organized as follows: Section 2 describes prior research efforts related to this work. Section 3 presents LCSM protocol specifications and its core operation. Section 4 illustrates simulation results and comparison to LOADng, AODV and LOADng-CTP. Finally, Section 5 concludes and summarizes the main contributions of this paper.

2 Ad Hoc Protocols for Power Line Communications

Several attempts to customize ad hoc routing protocols for power line communications are found in the literature.

Shucheng et al. introduces an on-demand multipath routing algorithm that tries to find maximally disjoint routes in large- scale networks with Master-Slave structure [2]. The protocol is able to build multiple routes using request/reply cycles. When the master requires a route to a given slave before knowing any routing information, it floods the RREQ message to the entire network. Several duplicates that traversed through different routes reach the destination as a result of flooding. Finally, the destination node picks up multiple disjoint routes from received RREQ packets and sends ROUTE REPLY (RREP) packets back to the source via the chosen routes. This scenario results in large network overhead and end-to-end delay.

Wei et al. [1] demonstrates a routing protocol based on AODV routing protocol. they modify two modules of the AODV, RREQ broadcasting mechanism and neighbor table manage- ment. The aim of these modifications is to reduce the overhead by reducing the hello packets.

In [3], Sivaneasan et al. proposes a routing algorithm based on non-overlapping clustering. It uses two-states Markov model for simulating the channel state during communication with the meters. In this protocol all meters have the role of relaying the DC message.

Zhenchao et al. [4] proposes a routing protocol based on overlapping clustering in order to establish different routes to reach the same meter which is useful at route failure condition. The DC selects cluster head that are responsible for delivering the data of neighboring meters to the DC.

Hong et al. [5] introduce a routing algorithm based on time slotted algorithm with random back off delay before transmission in order to reduce the collision.

In [6] Wenbing et al. proposes a routing protocol based on ant-colony algorithm which is described in [7]. There are two main tables that should be constructed; central routing table and pheromone routing table. Each child node can establish sub- routing table. Central node and child nodes need to establish their amplitude parameter list. Due to the response signal of child node, central node can set up sub-routing table one by one and update pheromone table. A greed stochastic adaptive searching method is also introduced in the ant colony optimization algorithm. One feature is that the establishment of the restricted candidate list (RCL) strategy. According to the amplitude parameters of the receiving signal among nodes, the RCL can be set up.

Clausen et al. [8] introduced LOADng routing protocol as a modified version of AODV protocol. LOADng outperforms AODV on packet delivery ratio and routing overhead. Jiazi et al. [9] improved the mechanism of RREQ compared to the basic LOADng in order to reduce the routing overhead. Furthermore, Jiazi et al. [10] introduce further modifications to LOADng and propose LOADng-CTP which is a class of collection-tree protocols and more suited to smart metering application. LOADng-CTP proves much better performance compared to LOADng and AODV.

Asymmetric characteristic of the power line channel has not been considered in the preceding protocols, where some nodes are subject to high line noise due to the household appliances [11]. Additionally, the preceding protocols assume that protocol algorithm will be programmed and executed at all nodes, whereas the majority of low-cost meters allow only for implementing very simple algorithms due to limited processing capability and on-board memory. LCSM takes the two aforementioned aspects into account. In the following section, specifications of LCSM is described.

3 LCSM Protocol Specification

The topology comprises Data Concentrator (DC) and several meters connected in a tree topology via low-tension power line grid. Figure 1 illustrates the topology of interest. The DC, as well as each meter, contains a power line modem that has a finite coverage range dependent on the maximum allowed transmitted power and receiver sensitivity. As shown, some meters could be accessed directly by the DC, while others are not reachable by the DC due to channel impairments, although they could be within the coverage range of intermediate meters.

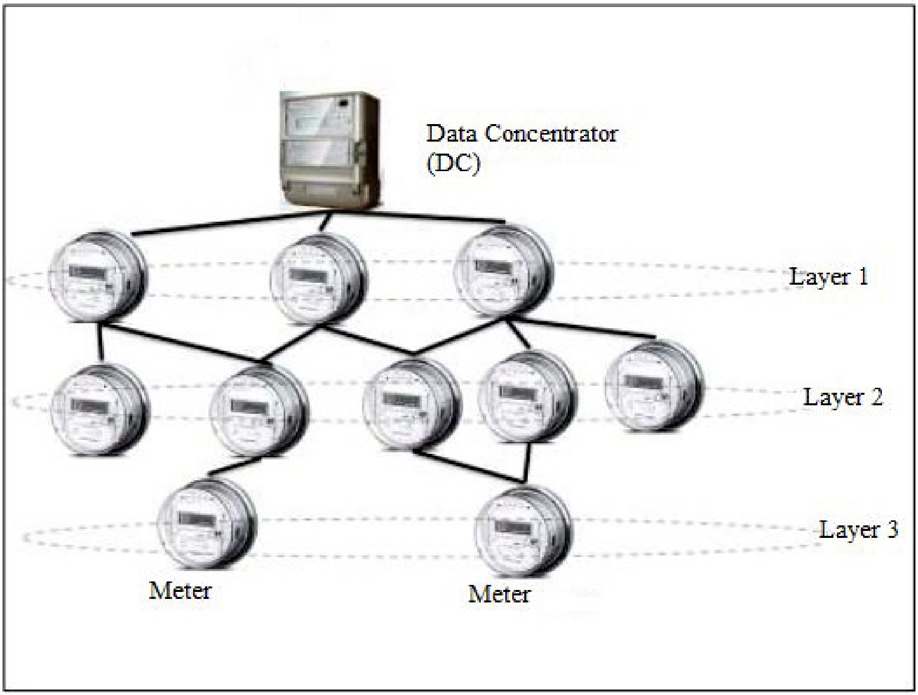


Fig. 1. Physical and Logical connection for smart metering system

3.1 Protocol Data Unit

LCSM utilizes two types of packets: Command packet and response packet. Command packets are the packets being sent from DC towards the meters and the response packets are those being sent from the meters towards the DC. Standard TLV (Type-Length-Value) packet format is used, as shown in Figure 2. The LCSM protocol utilizes source routing, so that the packet source-destination route is embedded within the packet body. Command packets are initiated from DC, while response packets are originated from meters. After topology discovery takes place, only the DC contains the complete visibility on network topology. Therefore, in command packets, the ID of each node along the route is included in order within the packet body. On the other hand, response packets contains only source, parent and final destination.

The PDUs used by LCSM are described as follows:

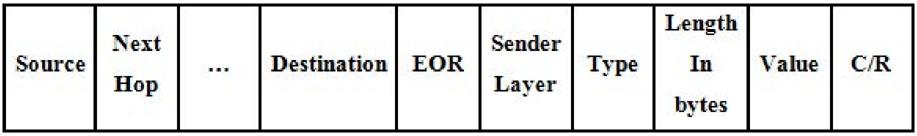
Neighbor Request (NREQ) - Neighbor Response (NRES): The purpose of NREQ packet is to explore who hears the DCs request. This is a broadcast packet. The response to this packet is NRES; the meter response with meter ID and the received SNR.

Layer/Parent Stamp (LPSTAMP) - Layer/Parent Acknowledgement (LPACK): LPSTAMP is used to inform the meter its parent-layer information. The response to this packet is LPACK.

Get Your Neighbors (GETN) - Neighbor List Reporting (NLREP): GETN packet is used to let certain parent reports its neighbors. The meter responds with NLREP packet that reports the list of which nodes are accessible by this specific parent.

Read Request (RREQ) - Read Response (RRES): After finishing the topology discovery cycle, the DC starts collecting the data (readings) using RREQ packet. The response to this packet is RRES packet.

Deaf Reading (DFREAD): DFREAD packet is used by the deaf node (the node that doesn't receive any request during certain pre-assigned time period is classified as DEAF) to broadcast its reading, which will be hopefully heard by adjacent node(s).



Source	Next Hop	...	Destination	EOR	Sender Layer	Type	Length In bytes	Value	C/R

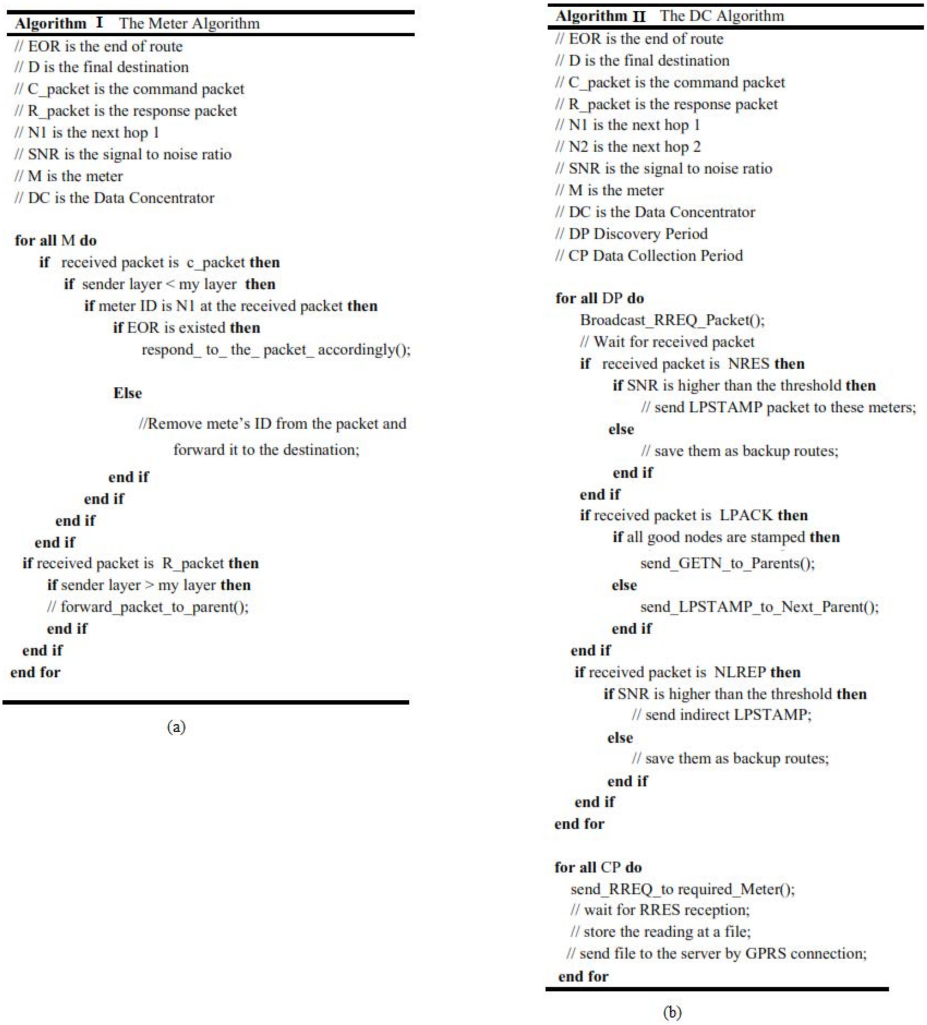
Fig. 2. Standard TLV packet format

Figures 3a and b illustrate the pseudo codes of LCSM algorithms for both meters and DC, respectively.

3.2 Protocol Routing Mechanism

Upon receiving a Command packet, the meter checks the node ID right after the source address. If it doesn't match the meter self ID, it ignores it. If it matches the meter ID, the meter checks whether it is the final node en-route (end of the route field - EOR), which represents the final packet destination. If it is, the meter responds according to packet type. If the meter ID doesn't lie at EOR field, this means that the meter lies within the source-destination route, and should act as a relaying node. Therefore, the meter relays the packet as it is after removing its ID field from the routing chain.

In Response packets, the meter checks the field representing parent ID, and if it matches self ID of the meter, it relays the packet to its parent by replacing the parent field with its own parent. (put Next Hop as P), and keep source and final destination the same. Whenever a meter receives a broadcast message from a deaf meter, it keeps the deaf meter ID, together with its reading. The meter reports deaf meter reading the next time its own reading is requested by the DC. According to this arrangement, the proper routing of packet only requires the knowledge of the node parent. Total routing matrix exists only at DC.



```
Algorithm I  The Meter Algorithm
// EOR is the end of route
// D is the final destination
// C_packet is the command packet
// R_packet is the response packet
// N1 is the next hop 1
// SNR is the signal to noise ratio
// M is the meter
// DC is the Data Concentrator

for all M do
    if  received packet is  c_packet then
        if  sender layer < my layer  then
            if meter ID is N1 at the received packet then
                if EOR is existed then
                    respond_ to_ the_ packet_ accordingly();

                Else
                    //Remove mete's ID from the packet and
                        forward it to the destination;
                end if
            end if
        end if
    end if
  if received packet is  R_packet then
      if sender layer > my layer then
      // forward_packet_to_parent();
      end if
  end if
end for
```

(a)

```
Algorithm II  The DC Algorithm
// EOR is the end of route
// D is the final destination
// C_packet is the command packet
// R_packet is the response packet
// N1 is the next hop 1
// N2 is the next hop 2
// SNR is the signal to noise ratio
// M is the meter
// DC is the Data Concentrator
// DP Discovery Period
// CP Data Collection Period

for all DP do
    Broadcast_RREQ_Packet();
    // Wait for received packet
    if  received packet is  NRES then
        if SNR is higher than the threshold then
            // send LPSTAMP packet to these meters;
        else
            // save them as backup routes;
        end if
    end if
    if received packet is  LPACK then
        if all good nodes are stamped then
            send_GETN_to_Parents();
        else
            send_LPSTAMP_to_Next_Parent();
        end if
   end if
   if received packet is  NLREP then
       if SNR is higher than the threshold then
           // send indirect LPSTAMP;
       else
           // save them as backup routes;
       end if
   end if
end for

for all CP do
   send_RREQ_to required_Meter();
   // wait for RRES reception;
   // store the reading at a file;
  // send file to the server by GPRS connection;
end for
```

(b)

Fig. 3. (a) Meter algorithm. (b) DC algorithm

4 Simulation Results and Performance Analysis

4.1 Simulation Environment

The LCSM protocol is simulated and evaluated by means of OPNET14.5 network simulator. Simulations are performed using number of nodes ranging from 50 to 500. The network is subject to multipoint-to-point (MP2P) traffic with all nodes generating traffic towards the Data Concentrator. Models of physical and Medium Access Control (MAC) layers of power-line modems are modeled using bus topology and an adaptive connectivity matrix. The purpose of the connectivity matrix, which is N x N (N is the number of nodes connected to power line) is to simulate whether a specific logical link between two nodes exist or not. In this way, PHY and MAC layers of the power line channel are modeled to allow for the application of LCSM at the network layer.

4.2 Simulation Parameters

The simulation parameters are summarized in Table 1. The power line channel could be considered physically as bus and logically as tree topology. Layer 1, 2 and 3 represent the tier at which the meters are located with respect to the Data Concentrator.

Table 1. Simulation Parameters

Parameter	Value
Number of Nodes	50-500
Simulation Time	100 seconds
Topology	Physically Bus, Logically Tree
MAC type	CSMA-CD
Slot time	0.214 second
Data rate	2400 bps
Preambe length	0 (no preamble)
Channel propagation delay	5.5 e^{-6} second
Type of service	Bursty traffic source
Burst duration	80 seconds
Burst period	5 seconds
Traffic type	Multi-Point-to-Point (MP2P)

4.3 Simulation Results

Comparison to AODV and LOADng

First, LCSM routing protocol is compared with LOADng [12] and AODV [13]. Although both protocols are originally designed for mesh network topology, the rationale behind comparing their performance to LCSM, which is a collection-tree protocol, is to highlight the expected enhancement resulting from using customized protocol for the smart metering case, which is by nature a tree topology. The results of LOADng and AODV are extracted from [12]. Figures 4 shows the simulation results for LOADng, AODV and LCSM routing protocols.

It is observed from Figure 4a that the overhead of LCSM is much lower than that of LOADng and AODV. The difference in overhead bytes is considerably higher at higher number of nodes. This is due to the large number of RREQ, RREP and RREP-ACK packets used in AODV for discovering the topology [13], [14]. At LCSM protocol, the parents are responsible for a lot of children which leads to reducing the overhead. It is important to study the end-to-end delay as only one (DC) is responsible for collecting data from around 400 meters. So, it is required to have a routing protocol with a controlled end-to-end delay, especially when dealing with time-critical events like load disconnect and fault isolation. As shown Figure 4b, LCSM routing protocol provides much lower end-to-end delay than LOADng and AODV. The variation of end-to-end delay at LCSM protocol is very small at large number of nodes.

As shown in Figure 4c , LCSM routing protocol introduces a delivery ratio which is very close to 100% regardless of the number of nodes.LOADng initiates route discovery for every router (network-wide broadcast) leads to a high number of collisions on the media, and thus a lower data delivery ratio, especially for larger number of nodes [10]. This is also applicable to AODV.

Figure 4d shows the topology discovery time against the number of nodes. It shows a considerable increase in discovery time after 200 nodes.

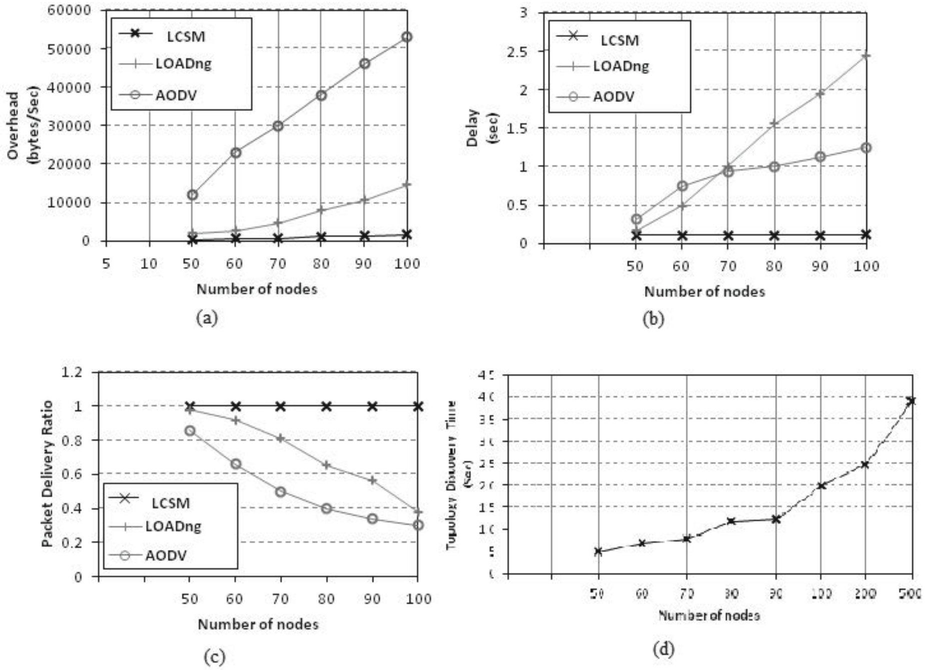


Fig. 4. (a) Routing overhead. (b) End-to-End delay. (c) Packet delivery ratio. (d) Topology discovery time at 2400 bps.

Memory Requirements for LCSM, AODV and LOADng:

For LOADng and AODV the memory requirements to store the routing table depends on the size of the network, the network topology and the number of traffic flows in the network. The contents of the routing table for LOADng protocol are:

(R_dest_addr, R_next_addr, R_metric, R_metric_type, R_hop_count,R_seq_num, R_bidirectional, R_local_iface_addr, R_valid_time) [8].

In LCSM protocol the only required entry to be stored at the meter is the parent address. The other routing information is contained at the message body. The overall matrix describing the topology is only stored at the DC.

Comparison to LOADng-CTP

Second, LCSM routing protocol is compared with LOADng-CTP [10]. Both protocols are collection-tree oriented, so both are optimized for smart metering application. As shown in Figure 5a, the number of bytes sent during the topology discovery process in both protocols are almost the same till reaching 200 nodes. At increased number of meters, the difference become larger and LOADng-CTP offers lower overhead. This is clear at 500 nodes. This performance is justified by the fact that the number of packets used for the topology discovery process in LOADng-CTP is less than the number of packets required by LCSM protocol, while the LCSM packet size is less than LOADng-CTP packet size. For this reason, the difference becomes obvious at higher number of nodes.

Another approach for evaluating the network overhead is based on the number of packets required to fully explore the topology. Figure 5b illustrates the value of this parameter against the number of nodes for both LCSM and LOADng-CTP. The similarity between the two protocols in terms of the number of packets required for topology exploration is obvious. However, as LCSM uses source routing,it is expected that with increasing number of nodes, the network depth (the maximum number of hops required to reach all nodes) increases, and therefore the average packet length will increase. This explains the fast increase in overhead bytes at LCSM compared to LOADng-CTP with increasing number of nodes, as illustrated in Figure 5b.

As mentioned previously, the data rate affects directly the end-to-end delay. Thus, it is predicted that the end-to-end delay for the packets sent at 2.4 kbps will be much greater than the packets sent at 11 Mbps as shown in Figures 5c. However, as the simulation results in [10] was performed at 11 Mbps data rate, it is required to evaluate the end-to-end delay of LCSM at the same data rate. Figure 5d illustrates the delay of both protocols when both are operating at 11 Mbps. It is clear that LCSM introduces smaller delay than LOADng-CTP at the range of nodes considered.

Figure 5e illustrates the packet delivery ratio of both protocols, and indicates that both protocols are identical and have a packet delivery ratio very close to 100%. This result is reasonable as the initiation of route discovery is made only for single destination, thus resulting in minimum number of collisions.

Memory Requirement for LOADng-CTP

For LOADng-CTP, only the route to the root is needed, and therefore one routing entry to the DC is required. This entry is defined by: (R_dest_addr, R_next_addr, R_metric, R_metric_type, R_hop_count,R_seq_num, R_bidirectional, R_local_iface_addr, R_valid_time) [8]

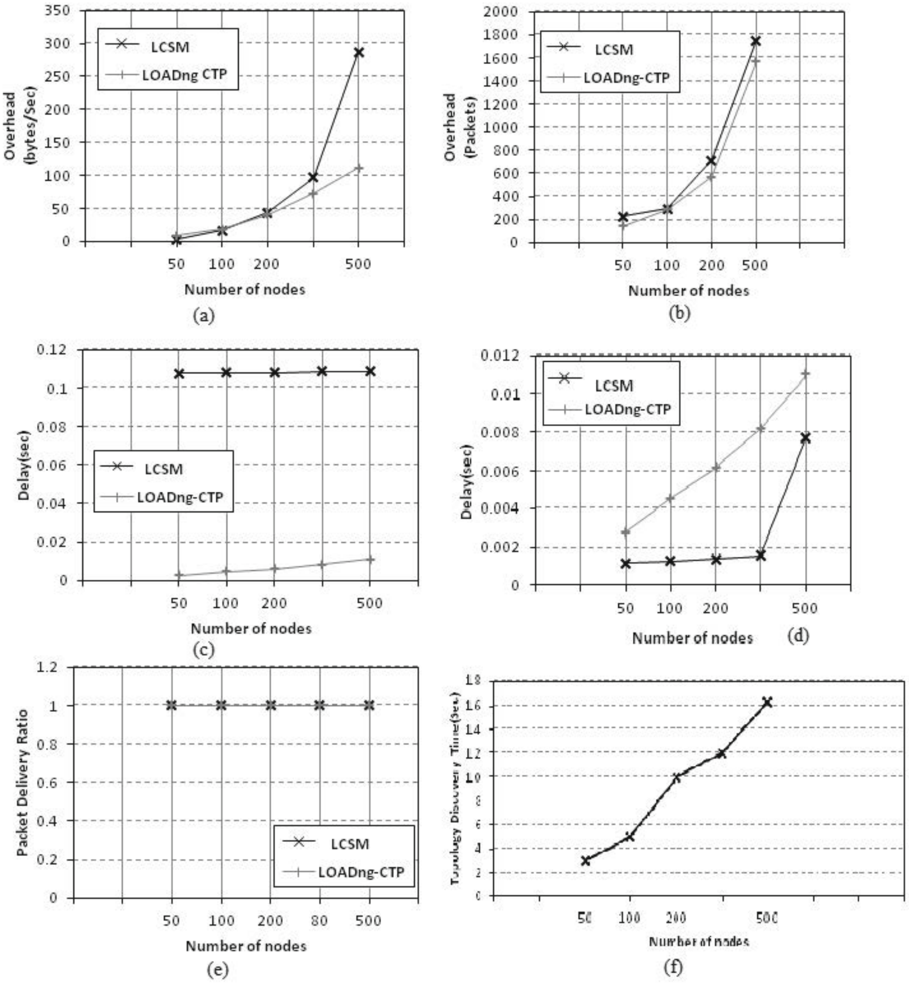


Fig. 5. (a) Routing Overhead for LCSM and LOADng-CTP(in bytes). (b) Routing Overhead for LCSM and LOADng-CTP(in packets). (c) End-To-End delay for LCSM (at data rate=2.4 kbps) and LOADng-CTP(at data rate=11 Mbps). (d) End-To-End delay for LCSM (at data rate=11 Mbps) and LOADng-CTP(at data rate=11 Mbps). (e) Packet Delivery Ratio for LCSM and LOADng-CTP. (f) Topology Discovery Time for LCSM at 11 Mbps.

The routing table entry for LOADng-CTP is much smaller than LOADng but it is higher than LCSM.

Figure 5f illustrates the topology discovery time for LCSM protocol with data rate 11 Mbps.

5 Conclusion

A low-complexity ad hoc routing protocol for smart metering over power line (LCSM) is proposed. A comparative analysis between the proposed protocol and AODV, LOADng and LOADng-CTP routing protocols is demonstrated. The simulation results show that LCSM routing protocol has considerably lower routing overhead compared to AODV and LOADng, especially at high number of nodes. It is also shown that the End-To-End delay of LCSM is lower than both LOADng and AODV, as the later are designed for mesh networks, while LCSM is a collection-tree oriented protocol. Comparison between LCSM and LOADng-CTP shows that the routing overhead is almost similar (LOADng-CTP is slightly better), the packet delivery ratio are almost the same (very close to 100%) and LCSM offers considerable lower end-to-end delay when running the simulation with the same data rate (11Mbps). Furthermore, algorithm complexity at the meter side when using LCSM is considerably reduced.

Table 2 summarizes the comparison between LCSM, LOADng-CTP, LOADng-CTP and AODV.

Table 2. Comparison between LCSM, LOADng-CTP, LOADng and AODV protocols

Comparison parameter	LCSM	LOADng-CTP	LOADng	AODV
Routing Overhead	Low	Low	High	High
End-to-End Delay	Low	Medium	High	High
Packet Delivery Ratio	High	High	Medium	Medium
Packet Format	Ethernet Format	AODV Format	AODV Format	AODV Format
Memory Requirement	Only parent should be saved	Complete routing table should be saved	Complete routing table should be saved	Complete routing table should be saved

References

1. Shucheng, L., Shumin, C., Xueli, D., Cuizhi, Z., Yuanxin, X.: A broadcasting algorithm of multipath routing in narrowband power line communication networks. In: Proceedings of the 3rd International Conference on Communication Software and Networks (ICCSN), pp. 467–471. IEEE (2011)

2. Wei, G., Wenguang, J., Hao, L.: An improved routing protocol for power-line network based on AODV. In: Proceedings of the 11th International Symposium on Communications and Information Technologies (ISCIT), pp. 233–237. IEEE (2011)
3. Sivaneasan, B., So, P., Gunawan, E.: A Simple Routing Protocol for PLC-based AMR Systems. In: Proceedings of the TENCON Conference, pp. 1–5. IEEE (2009)
4. Zhenchao, W., YiJin, W., Jing, W.: Overlapping Clustering Routing Algorithm Based on L-PLC Meter Reading System. In: Proceedings of the International Conference on Automation and Logistics (ICAL), pp. 1350–1355. IEEE (2009)
5. Hong, L., Huang, D.: A Time Slotted Multiple Access Control protocol with real time quality for Low Voltage Power-line Carrier Network. In: Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, pp. 136–140. IEEE (2010)
6. Wenbing, L., Yingli, L., Xiaowei, B., Yonghong, M., Yongquan, C.: Study on automatic relaying algorithm for PLC based on channel state. In: Proceedings of the 2nd International Conference on Communication Systems, Networks and Applications, pp. 81–85. IEEE (2010)
7. Dorigo, M., Vittorio, M., Alberto, C.: The Ant System: Optimization by a colony of cooperating agents. Proceedings of the Transactions on Systems, Man, and Cybernetics-Part B (26), I-B (1996)
8. Clausen, T., de Verdiere, A.C., Yi, J., Niktash, A., Igarashi, Y., Satoh, H.: The Lightweight On-demand Ad hoc Distance-vector Routing Protocol - Next Generation (LOADng). The Internet Engineering Task Force, Work in Progress, draft-clausen-lln-loadng-08). IETF (January 2013)
9. Yi, J., Clausen, T., Bas, A.: Smart Route Request for On-demand Route Discovery in Constrained Environments. In: Proceedings of the Wireless Information Technology and Systems (ICWITS), pp. 1–4. IEEE (2012)
10. Yi, J., Clausen, T., Bas, A.: Smart Route Request for On-demand Route Discovery in Constrained Environments. In: Proceedings of the Wireless Information Technology and Systems (ICWITS), pp. 1–4. IEEE (2012)
11. Murty, R., Padhye, J., Chandra, R., Chowdhury, A.R., Welsh, M.: Characterizing the End-to-End Performance of Indoor Powerline Networks, in Technical Report. Harvard University (2008)
12. Clausen, T., Yi, J., de Verdiere, A.C.: LOADng: Towards AODV Version 2. In: Vehicular Technology Conference (VTC Fall), pp. 1–5. IEEE (2012)
13. Perkins, C., Belding-Royer, E., Das, S.: Ad hoc On-Demand Distance Vector (AODV). Experimental RFC 3561 (July 2003)
14. Clausen, T., de Verdiere, A.C., Yi, J., Niktash, A., Igarashi, Y., Satoh, H.: Interoperability Report for the Lightweight On-demand Ad hoc Distance vector Routing Protocol - Next Generation (LOADng) draft-lavenu-lln-loadng-interoperability-report-04, in The Internet Engineering Task Force. IETF (December 2012)

Algorithms for Joint Coverage, Routing and Scheduling Problem in Wireless Sensor Networks

Seweryn Jagusiak and Jerzy Józefczyk

Insitute of Informatics, Wroclaw University of Technology, Wrocław, Poland

Abstract. A new decision making problem for wireless sensor networks is considered in the paper. A coverage of data sources, the routing of measured data as well as a scheduling of working periods of sensors are assumed as the decision to be made. All decisions as interconnected are determined jointly. The energy consumption and execution time of sensors are used as criteria. The corresponding combinatorial NP-hard optimization problem is formulated. To solve it, two heuristic algorithms are proposed. Some initial analitical evaluations of algorithms are presented as well as the result of computational experiments are given.

Keywords: scheduling, routing, coverage, wireless sensor network, heuristic algorithm, approximation.

1 Introduction

A progress in new optimization methods and algorithms as well as in corresponding computing tools makes it possible to investigate more complex problems which are closer to real world applications. It concerns operations research problems, in general, and combinatorial problems, in particular, where complex optimization problems being the combination of interrelated, known and separately developed sub-problems are intensively studied.

Wireless sensor networks (WSNs) are examples where such approach can be applied. A standard sensor network consists of a set of sensors, which are small electronic devices capable to collect data on certain phenomena from a defined area and then to process and transmit them, as well as a set of sinks (hubs, gates) intended for gathering the data and providing them to users. Measured phenomena are called hereinafter data sources (Fig. 1). Sensors usually work in the environment that makes impossible their constant maintenance, so, the proper energy management is the crucial task enabling the lifetime of WSNs. It does not concern sinks which unlike sensors are constantly supplied facilities. Measuring time is also important in some types of WSNs where a fast reaction to changes is needed.

Development of WSNs enables the significant extension of their applications beyond the original military usage. It is possible by advances in miniaturized mechatronic systems, and first of all, in a wireless communication. So, the contemporary researches on WSNs have interdisciplinary nature and belong not only

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_56, © Springer International Publishing Switzerland 2014

to computer science but also to meteorology, electrical engineering and telecommunication. In general, they consist in: designing of sensors, their deployment in an appropriate working environment, determination of coverage areas for sensors, collection and transmission of data, minimization of the energy consumption by sensors, ensuring the security of data during measurement, collection and transmission, guarantee of the quality of services, solving various particular issues connected with the mobility of sensors, e.g.[1].

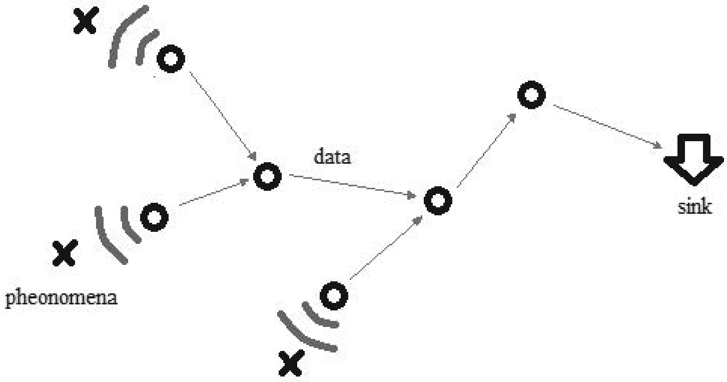


Fig. 1. Wireless sensor network scheme

The development of both stationary and mobile WSNs generates also interesting problems in the area of operations research, which need solutions. Management of WSNs, which leads to the extension of the execution time of sensors, is the main challenge. The most important research problems from this scope for stationary networks are: routing of data acquired by sensors to sinks possibly via other sensors treated as brokers in the transmission as well as coverage of sensing areas by sensors, e.g. [2] and [3]. These problems are solved separately in many works such as [4,5]. The coverage problem is particularly important. Energy efficient assignment sensors to data sources may increase their execution times. Otherwise, fast measurement times can be acquired by using high energy consuming sensors, which in a consequence, increases the total energy consumption of WSNs. In the literature, many different methods to formulate and solve a coverage problem are reported, e.g.: binary formulation of coverage [4], probabilistic approach [6]. The special case of the problem is also considered when sensors have to operate a finite number of measurement points rather than a planar area with the infinite number of points. Then, the coverage problem can be expressed using values of corresponding variables characterizing the problem, e.g. number of units of data that need to be acquired by sensors. Such an approach is used in this work. It is assumed that the coverage depends on the location of sensors and their type, due to the fact that sensors can have different sensing ranges, energy consumption or operation execution times. Much more attention has been paid in previous studies to routing problems. In order to minimize the energy, the multihop transmission is often used. It means that sensors serve as brokers in the transmission and just forward data to other sensors or sinks.

The scheduling is obviously connected with the data transmission. Each sensor can turn off their sensing device and receiver what brings important impact on the energy usage. On the other hand, a schedule can be determined when operation is made, i.e: when data are acquired from phenomena or when the transmission between brokers starts. Due to interconnection between routes and schedules, there is a need to consider these problems jointly. The routing problem in WSNs is often investigated together with the coverage problem, e.g., [6]. In some works, it is formulated and solved as a clustering problem [7]. There are also papers which solve joint problems with scheduling , i.e: [8,9].

All three decisions, i.e.: the scheduling, the routing of data and the coverage of sensing area by sensors are not yet investigated jointly. In this paper, two criteria i.e. the energy usage and the time execution are used to evaluate mentioned decisions.

The remaining text is organized as follows. The formulation of the problem for a single commodity data transmission and its analysis are given in Section 2. Section 3 presents heuristic solution algorithms which use the solutions of the shortest path problem. The results of numerical experiments assessing the algorithm are discussed in Section 4. Final remarks complete the paper.

2 Problem Formulation

Let us consider WSN with already deployed set of sensors $\mathbf{S} = \{1, 2, 3, ..., S\}$. The aim of the network is to acquire by sensors data from data sources belonging to set $\mathbf{P} = \{S + 1, S + 2, S + 3, ..., S + P\}$ and to send them towards deployed sinks from set $\mathbf{U} = \{P+S+1, P+S+2, P+S+3, ...P+S+U\}$. Data can be sent to sinks directly from the measuring sensor or using other sensors as brokers. Sensor can transmit at a time data to one element and turn off connection when it is not used.

We use the following additional notation:

$\mathbf{T} = \{1, 2, ..., t, ..., T\}$ - set of time periods in which data are acquired and transmitted to sinks, T time horizon,

$\mathbf{N} = \mathbf{P} \cup \mathbf{S} \cup \mathbf{U} = \{1, 2, 3, ..., N\}$- set of all elements,

e_{ijk} – execution time of acquirement or transmission of data originating in kth source from element i to j if respectively $i \in \mathbf{P}, j \in \mathbf{S}$ or $i \in \mathbf{S}$, $j \in \mathbf{S} \cup \mathbf{U}$, $\mathbf{e} = \{e_{ijk}\}, (i, j) \in \{(a, b) : a \in \mathbf{P} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{U}\}, k \in \mathbf{P}$,

c_{ijk} – energy consumption of acquirement or transmission of data originating in kth source from element i to j if respectively $i \in \mathbf{P}, j \in \mathbf{S}$ or $i \in \mathbf{S}$, $j \in \mathbf{S} \cup \mathbf{U}$, $\mathbf{c} = \{c_{ijk}\}, (i, j) \in \{(a, b) : a \in \mathbf{P} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{U}\}, k \in \mathbf{P}$,

The set x of decision variables contains binary variables x_{ijkt}, $(i, j) \in \{(a, b) : a \in \mathbf{P} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{U}\}, t \in \mathbf{T}, k \in \mathbf{P}$, where $x_{ijkt} = 1(0)$

if data acquirement or transmission from element i to j derived from source k begins in period t (otherwise).

The following constraints are imposed on x to ensure determining of feasible solutions:

$$\sum_{t \in \mathbf{T}} \sum_{j \in \mathbf{S}} x_{ijit} = 1 \qquad \forall_{i \in \mathbf{P}}, \tag{1}$$

$$\sum_{t \in \mathbf{T}} \sum_{j \in \mathbf{S} \cup \mathbf{U}} x_{ijkt} = \sum_{t \in \mathbf{T}} \sum_{j \in \mathbf{S} \cup \mathbf{P}} x_{jikt} \qquad \forall_{i \in \mathbf{S}, k \in \mathbf{P}}, \tag{2}$$

$$e_{ijk}(1 - x_{ijkt}) > \sum_{j' \in \mathbf{S} \cup \mathbf{U}} \sum_{k' \in \mathbf{P}} \sum_{t' \in \{t,\ldots,t-1+e_{ijk}\}} x_{ij'k't'} - x_{ijkt} \ \forall_{i \in \mathbf{S}, j \in \mathbf{S} \cup \mathbf{U}, k \in \mathbf{P}, t \in \mathbf{T}}, \tag{3}$$

$$x_{ijkt} \leq \sum_{j' \in \mathbf{S} \cup \{k\}} \sum_{t' \in \{1,..,t-1\}} x_{j'ikt'} \ \forall_{i \in \mathbf{S}, j \in \mathbf{S} \cup \mathbf{U}, k \in \mathbf{P}, t \in \mathbf{T}}, \tag{4}$$

$$x_{jikt}(t + e_{jik}) \leq h \qquad \forall_{i \in \mathbf{U}, j \in \mathbf{S}, t \in \mathbf{T}, k \in \mathbf{P}}. \tag{5}$$

Equation (1) ensures that all data sources are covered by sensors. Constraint (2) is responsible for the flow balance between the number of incoming and outcomming data for each sensor. Constraint (3), (4) ensure proper scheduling, i.e. sensor can transmit some data: once a time, only after receiving it. Last equation (5) is used to determine the execution time.

To evaluate the decision x, the joint criterion $Q(x)$ is proposed being a weighted sum of two subcriteria: total energy consumption$Q_e(x)$ and the execution time $Q_t(x)$, i.e.

$$Q(x) = \alpha Q_e(x) + (1 - \alpha) Q_t(x) =$$

$$\alpha \sum_{i \in \mathbf{S}} \sum_{t \in \mathbf{T}} \sum_{k \in \mathbf{P}} \Big(\sum_{j \in \mathbf{S} \cup \mathbf{U}} c_{ijk} x_{ijkt} + c_{kik} x_{kikt} \Big) + (1 - \alpha) \beta h \tag{6}$$

where coefficient $\alpha \in [0, 1]$ reflects the importance of individual subcriterion and cooeficient β constitutes the desirable proportion between the energy consumption and the execution time.

Finally, for given $\mathbf{P}, \mathbf{S}, \mathbf{U}, \mathbf{e}, \mathbf{c}$, the problem considered in the paper deals with the determination of x feasible with respect to (1) - (5) to minimize (6), i.e. $Q^* \triangleq min_x Q(x)$.

It is worth noting that for $\alpha = 1$, i.e. when the execution time is not taken into account, the problem becomes easy due to the fact that then each data source can be treated independently, and it is enough to transmit all data originating there to sinks via corresponding shortest paths. For $\alpha < 1$, problem is NP hard because the set cover problem can be reduced to it.

3 Algorithms and Analysis

We propose two heuristic algorithms to solve the problem formulated in Section 2. The first one, being the P-approximate algorithm, is based on the solution of the shortest path problem (SPP). The second algorithm uses additionally a specific approach for the determination of solutions.

We denote auxiliary decision variable $x' = \{x'_{ijk}\}, (i,j) \in \{(a,b) : a \in \mathbf{P} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{S} \vee a \in \mathbf{S} \wedge b \in \mathbf{U}\}, k \in \mathbf{P}$ which is responsible for routing and coverage determination. Its value corresponds to decision variable x_{ijkt} as follows: $x'_{ijk} = 1 \Leftrightarrow \exists_{t' \in \mathbf{T}} x_{ijkt'} = 1$.

Let us consider release dates (time when sensor is ready to perform transmission to sink or to receive data) for sensors and sinks as auxiliary variable $t^0 = \{t^0_n\}, n \in \mathbf{S} \cup \mathbf{U}$. Their values can be calculated iteratively using variable x'_{ijk}

$$t^0_n = \begin{cases} \max\limits_{i \in \mathbf{P} \cup \mathbf{S} : \exists_{k \in \mathbf{P}} x'_{ink} = 1} (t^0_i + \sum\limits_{k \in \mathbf{k}} x'_{ink} e_{ink}), & n \in \mathbf{S} \cup \mathbf{U} \\ 1, & n \in \mathbf{P}. \end{cases} \tag{7}$$

The values of t^0_n and x'_{ijk} enable us to determine decision x. Before the presentation of proposed algorithms, it is convenient to give the auxiliary procedure, called Scheduling Procedure, used by them. As t^0 can be calculated in linear time, the complexity of Scheduling Procedure is $O(N^3)$.

Scheduling Procedure

```
Data: $t^0, x', \mathbf{e}$
Result: $x$
$x_{ijkt} = 0$ (for all indexes)
for all pairs $(i,j) \in (\mathbf{P} \cup \mathbf{S}) \times \mathbf{S} \cup \mathbf{S} \times \mathbf{U}$ do
    Create list $Q$ of sources $k$ for which $x'_{ijk} = 1$, sorted descending according to $e_{ijk}$
    $t_{sum} = t^0_i$.
    for all $k \in Q$ do
        $x_{ijkt_{sum}} = 1$
        $t_{sum} = t_{sum} + e_{ijk}$
    end for
end for
```

3.1 *P*-Approximation Algorithm

This algorithm referred to as Algorithm 1 works in two phases. At first, the algorithm seeks for the shortest paths from each data source point $SP(p)$ to any sink what is responsible for the determination of routes and coverages. Then, the schedule is calculated using Scheduling Procedure.

While using Dijkstra's algorithm for determining the shortest paths, the complexity of Algorithm 1 is $O(N^3)$.

As it has been mentioned, Algorithm 1 returns the optimal solution for $\alpha = 1$.

For $\alpha = 1$ proposed algorithm returns optimal solution. Let's observe that for $\alpha = 0$ the optimal value of (6) is greater or equal to the longest time path, i.e. $Q^* \leq max_{p \in \mathbf{P}} SP(p)$. In the worst case, Algorithm 1 would select the same path

Algorithm 1. P-approximation algorithm

```
Data: $\mathbf{e}, \mathbf{c}, \mathbf{P}, \mathbf{S}, \mathbf{U}$
Result: $x$
$x'_{ijk} = 0$ (for all indexes)
for all $p \in \mathbf{P}$ do
    Create undirected graph $G = (V, L)$ with edge costs $d_l$ $(l \in L)$, where:
    $V = \{p, r\} \cup \mathbf{S} \cup \mathbf{U}$ ($r$ - extra vertex),
    $d_l = \alpha c_{ijp} + (1 - \alpha)\beta e_{ijp}, l = (i, j) \in (\{p\} \times \mathbf{S}) \cup (\mathbf{S} \times (\mathbf{S} \cup \mathbf{U}))$,
    $d_l = 0, l \in \mathbf{S} \times \{r\}$,
    $d_l = \infty$ for other edges.
    Solve SPP problem from vertex $p$ to vertex $r$ with edges' costs $d_l$ to obtain $SP(p)$.
    Set $x'_{i^* j^* p} = 1$ for $l^* = (i^*, j^*)$, being the result of solving SPP.
end for
Determine $x$ by Scheduling Procedure using $x'$.
```

for all sources, so $Q^1 \leq P \cdot max_{p \in \mathbf{P}} SP(p)$, where Q^1 is the value of (6) returned by Algorithm 1. In a consequence,

$$Q^1 \leq P \cdot max_{p \in \mathbf{P}} SP(p) \leq PQ^*. \tag{8}$$

In general when $0 < \alpha < 1$, each value of SP solution consists of the energy part with factor α and the time part with factor $1 - \alpha$. Consequently, we know that the latter part $Q_t^1(x)$ is calculated with P approximation. Let us observe that improving (minimizing) the value of criterion Q^1 as a whole would decrease the time part and increase the energy part, so Q_e^1 is not greater than Q_e^*. So, we get

$$Q^1 = \alpha Q_e^1 + (1 - \alpha) Q_t^1 \leq \alpha Q_e^* + (1 - \alpha) P Q_t^* = PQ^* - \alpha(P - 1) Q_e^* \leq PQ^* \tag{9}$$

what proves the approximation ratio of Algorithm 1.

Additionally, we can propose that the tighter approximation of Q_t^1 which, however, is the result of Algorithm 1. Namely,

$$PQ^* - \alpha(P - 1) Q_e^* \leq PQ^* - \alpha(P - 1) Q_e^1 = (P - \frac{P - 1}{(1 - \alpha) Q_t^1}) Q^*. \tag{10}$$

3.2 Constructive Algorithm

This algorithm is similar to the previous one but uses its partial solutions. The data sources are sorted descendingly according to $SP(p)$. Then, the values of paths are calculated again according to the order with different edges costs in each step.

The time complexity of Algorithm 2 is not higher than Algorithm 1. However, Algorithm 2 works about two times longer then Algorithm 1. Algorithm 2 does not preserve optimality for $\alpha = 1$ but preserves P approximation ratio like Algorithm 1.

At the beginning, the order of data sources is calculated starting from the most costly path determined by the algorithm solving SPP. The first path with

Algorithm 2. Constructive algorithm

```
Data: $\mathbf{e}, \mathbf{c}, \mathbf{P}, \mathbf{S}, \mathbf{U}$
Result: $x$
Auxiliary variable: $F$ as list of pairs $((p_1, SP(p_1)), ..., (p_P, SP(p_P)))$
$x'_{ijk} = 0$ (for all indexes)
for all $p \in \mathbf{P}$ do
    Create undirected graph $G = (V, L)$ with edge cost $d_l$ $(l \in L)$ where:
    $V = \{p, r\} \cup \mathbf{S} \cup \mathbf{U}$ ($r$ - extra vertex)
    $d_l = \alpha c_{ijp} + (1-\alpha)\beta e_{ijp}, l = (i,j) \in (\{p\} \times \mathbf{S}) \cup (\mathbf{S} \times (\mathbf{S} \cup \mathbf{U}))$,
    $d_l = 0, l \in \mathbf{S} \times \{r\}$,
    $d_l = \infty$ for other edges.
    Solve SPP from vertex $p$ to vertex $r$ with edges' costs $d_l$ $(SP(p))$.
    Add pair $(p, SP(p))$ to $F$.
end for
     Sort $F$ descendingly according to second element of pairs ($SP$ value).
for all  $p \in F$ do
    Create undirected graph $G = (V, L)$ with edge cost $d_l$ $(l \in L)$, where:
    $V = \{p, r\} \cup \mathbf{S} \cup \mathbf{U}$ ($r$ - extra vertex),
    $d_l = \alpha c_{ijp} + (1-\alpha)\beta e_{ijp} + \sum_{k \in \mathbf{P}} (\alpha c_{ijk} + (1-\alpha)\beta e_{ijk}) x'_{ijk}$,
    $l = (i,j) \in (\{p\} \times \mathbf{S}) \cup (\mathbf{S} \times (\mathbf{S} \cup \mathbf{U}))$,
    $d_l = 0, l \in \mathbf{S} \times \{r\}$,
    $d_l = \infty$ for other edges.
    Solve SPP from vertex $p$ to vertex $r$ with edges' cost $d_l$.
end for
Determine $x$ by Scheduling Procedure on the basis of $x'$.
```

value $max_{p \in \mathbf{P}} SP(p)$ is included in the solution. Let us observe that the next path is added with cost no bigger then previous one because the consecutive cost from the list is less than previous one , and in the worst case Algorithm 2 can choose one of already selected paths, which is not greater than $max_{p \in \mathbf{P}} SP(p)$. Hence, the total cost is not greater than $P \cdot max_{p \in \mathbf{P}} SP(p)$, as for Algorithm 1.

4 Computational Experiments

In order to evaluate the quality of the heuristic solution algorithms proposed, the preliminary computational experiments were performed. All elements were randomly deployed according to the uniform distribution on the 100x100 units square area. Elements of matrices **c** and **e** were calculated proportional to the distances among WSN elements, and the latter ones were normalized to not exceed value 5 as well as rounded up to the nearest integer number. It was assumed that $S = 6, U = 2, \beta = 10^7$.

The quality of results generated by the heuristic algorithms was assessed for different values of parameters α and P. The examples of results are presented in Table 1 where corresponding values are averages of 10 independent runs of each algorithm. Solutions generated by the heuristic algorithms were evaluated with reference to the optimal solutions obtained by solver GLPK (http://www.gnu.org/s/glpk), using the performance indecies $\frac{Q^2}{Q^*}$, $\frac{T^*}{T^1}$, $\frac{Q^2}{Q^*}$, $\frac{T^*}{T^2}$ where Q^2 is the value of (6) calculated by Algorithm 2 and T^1, T^2 are runtimes of Algorithm 1, Algorithm 2, respectively.

The following main conclusions result from the experiments conducted:

a. All values of (6) determined by both algorithm are consistent with approximation ratio P.
b. As it was expected, Algorithm 1 gives worse results than Algorithm 2 as the latter one improves the results obtained by the former one.
c. Algorithm 1 is about 2 times faster than the other heuristic. However, both of them work in the reasonable time unlike the exact algorithm which times grow very fast when increasing the size of the problem.
d. Both heuristic algorithms find better solutions for smaller values of α.
e. Algorithm 2 gives better results for greater number of data sources P.

Table 1. Results of computional experiments evaluating Algorithm 1 and 2 for different P and α

P	α	$\frac{Q^1}{Q^*}$	$\frac{T^*}{T^1}$	$\frac{Q^2}{Q^*}$	$\frac{T^*}{T^2}$	P	α	$\frac{Q^1}{Q^*}$	$\frac{T^*}{T^1}$	$\frac{Q^2}{Q^*}$	$\frac{T^*}{T^2}$
5	0,75	1,98	7,79	1,29	3,93	5	0,25	1,15	11,78	1,09	6,13
6	0,75	1,77	18,57	1,04	11,28	6	0,25	1,32	22,15	1,10	14,76
7	0,75	1,56	27,98	1,08	17,95	7	0,25	1,68	32,45	1,24	18,54
8	0,75	1,48	33,72	1,06	19,25	8	0,25	1,75	39,45	1,25	23,05
9	0,75	1,60	68,47	1,09	38,33	9	0,25	1,60	58,30	1,13	37,69
10	0,75	1,74	52,45	1,16	28,76	10	0,25	2,01	53,63	1,29	28,44
5	0,5	1,29	10,04	1,11	6,96	5	0	2,76	7,357	1,46	4,25
6	0,5	1,34	12,71	1,05	7,81	6	0	2,77	15,62	1,07	8,16
7	0,5	1,66	23,90	1,23	14,12	7	0	2,21	23,69	1,17	14,79
8	0,5	2,18	34,30	1,39	19,01	8	0	2,46	27,39	1,06	16,38
9	0,5	2,44	41,03	1,25	23,45	9	0	2,29	73,24	1,50	39,66
10	0,5	2,36	96,46	1,15	50,05	10	0	2,57	70,62	1,26	38,60

5 Final Remarks

The selected version of the joint coverage, scheduling and routing problem for wireless sensor networks is considered in the paper. Two heuristic algorithms were proposed which based on the solution of the shortest path problem in graphs. The analysis of these algorithm showed their approximation properties. The numerical experiments assessing the quality of the algorithm have been also performed. It turned out that for all instances tested it is possible to obtain solutions at most 1.5 times worse than the optimal ones.

During further works we will be concerned with the following research directions:

a. Consideration of new versions of the problem where decisions are made in decentralized structure and online.
b. Integration of other criteria, for example with energy or time part moved to the constraints.
c. Seeking for more efficient approximation schemes.
d. Taking into consideration additional constraints and assumptions, e.g. capacity of sensors,
e. Implementation of the algorithms proposed and testing them in laboratory environment.

References

1. Kulkarini, R.V., Fox Andrster, A., Venayagamoorthy, G.K.: Computational intelligence in wireless sensor networks: A survey. Communications Surveys Tutorials 13(1), 68–96 (2011)
2. Bin, W., Wenxin, L., Liu, L.: A survey of energy conservation, routing and coverage in wireless sensor networks. In: Zhong, N., Callaghan, V., Ghorbani, A.A., Hu, B. (eds.) AMT 2011. LNCS, vol. 6890, pp. 59–70. Springer, Heidelberg (2011)
3. Chuan, Z., Chunlin, Z., Lei, S., Guangjie, H.: A survey on coverage and connectivity issues in wireless sensor networks. Journal of Network and Computer Applications 35(2), 619–632 (2012)
4. Zou, L., Lu, M., Xiong, Z.: A distributed algorithm for the dead end problem of location based routing in sensor networks. IEEE Transactions on Vehicular Technology 54(4), 1509–1522 (2005)
5. Shankarananda, B.M., Saxena, A.: Energy efficient localized routing algorithm for wireless sensor networks. In: Proceedings of 3rd International Conference on Electronics Computer Technology (ICECT), pp. 72–75 (2011)
6. Guney, E., Aras, N., Altinel, I.K., Ersoy, C.: Efficient solution techniques for the integrated coverage, sink location and routing problem in wireless sensor networks. Computers & Operations Research 39(7), 1530–1539 (2012)
7. Abbasi, A.A., Younis, M.: A survey on clustering algorithms for wireless sensor networks. Computer Communications 30(14-15), 2826–2841 (2007)
8. Mo, W., Qiao, D., Wang, Z.: Mostly-sleeping wireless sensor networks: connectivity, k-coverage, and alpha-lifetime. In: Proceedings of the 43rd Annual Allerton Conference on Communication, Control, and Computing (2005)
9. Deng, J., Han, Y.S., Heinzelman, W.B., Varashney, P.K.: Scheduling sleeping nodes in high density cluster-based sensor networks. Mobi. Netw. Appl. 10(6), 825–835 (2005)
10. Chuan, Z., Chunlin, Z., Lei, S., Guangjie, H.: A survey on coverage and connectivity issues in wireless sensor networks. Journal of Network and Computer Applications 35(2), 619–632 (2012)
11. Golden, B.: Shortest-Path Algorithms: A Comparison. Operations Research 24(6), 1164–1168 (1976)
12. Jagusiak, S., Józefczk, J.: An algorithm for joint location, coverage and routing in wireless sensor networks. In: XII BOS Conference, Polish Operational and Systems Research Society, Warsaw (2012)

Cyclic Scheduling of Multimodal Concurrently Flowing Processes

Grzegorz Bocewicz[1], Robert Wójcik[2], and Zbigniew Banaszak[3]

[1] Koszalin University of Technology, Dept. of Computer Science and Management,
ul. Sniadeckich 2, 75-453 Koszalin, Poland
bocewicz@ie.tu.koszalin.pl

[2] Institute of Computer Engineering, Control and Robotics,
Wrocław University of Technology, Wrocław, Poland
robert.wojcik@pwr.wroc.pl

[3] Warsaw University of Technology, Dept. of Business Informatics,
Warsaw, Poland
z.banaszak@wz.pw.edu.pl

Abstract. The problem of cyclic scheduling of multimodal concurrent processes (MCPs) is considered. Processes composed of sub-sequences of local cyclic processes (LCPs) are treated as multimodal processes, e.g. in case local processes encompass the subway lines network the relevant multimodal processes can be seen as passengers traveling itineraries this network. Since LCPs network implies a MCPs behavior, the following fundamental questions arise: Does the given LCPs network guarantee an assumed cyclic schedule of the MCPs at hand? Does there exist the LCPs network such that an assumed cyclic schedule of MCPs can be achieved? In that context our contribution is to propose a declarative framework allowing one to take into account both direct and reverse problems formulation.

Keywords: cyclic scheduling, multimodal processes.

1 Introduction

Cyclic scheduling problems arise in different application domains (such as manufacturing, time-sharing of processors in embedded systems) as well as service domains (covering such areas as workforce scheduling, timetabling, and reservations) [1], [2], [4], [5], [6], [12], [14]. Subway or train [14] traffic can be considered as an example of such kind of systems as well as flow shops, job shops, assembly systems, and, in general, to any discrete event system which transforms raw material and/or components into products and/or components [7], [8]. Among the manufacturing systems the ones providing constantly the required mixture of various products, through the realizing so called cyclic manufacturing policy focus a special attention. In such systems each product is performed along the unique sequence of operations on dedicated machines, defined by individual technological route, and different for different products. The aim is to find the

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_57, © Springer International Publishing Switzerland 2014

cyclic schedule of minimal cycle length. The considered cyclic job shop scheduling problem, is strongly NP-hard [12].

Many models and methods have been proposed to solve the cyclic scheduling problem [7]. Among them, the mathematical programming approach (usually IP and MIP), max-plus algebra [8], constraint logic programming [2], [3], evolutionary algorithms [10] and Petri net [12] frameworks belong to the more frequently used. Most of them are oriented at finding of a minimal cycle or maximal throughput while assuming deadlock-free processes flow. The approaches trying to estimate the cycle time from cyclic processes structure and the synchronization mechanism employed (i.e. rendezvous or mutual exclusion) are quite unique [1], [2], [3], [8]. In that context, our approach to cyclic scheduling of multiproduct manufacturing within FMS environment, especially to cyclic scheduling of associated AGVS, employing declarative modeling approach while implementing a concept of concurrently flowing cyclic processes (SCCP) can be seen as continuation of our former work [1], [2], [3], [8].

The approach proposed permits to determine the model for the assessment of the impact of the structure of LCPs on the MCPs performance indices as well as provides the declarative framework allowing one to take into account both direct and reverse problems formulation of the MCPs scheduling. That can be seen as extension of our former work [1], [2], [3], [8], while aimed at pipeline-like flows of local as well as multimodal processes. So, the papers objective is to provide the conditions useful in the course of tasks routing and scheduling in systems composed of cyclic processes interacting each other through mutual exclusion protocol.

The rest of the paper is organized as follows: Section 2 introduces to the concept of multimodal processes and states their scheduling problem. The Section 3 introduces to multimodal scheduling problem. In Section 5, a declarative modeling framework is implemented to illustrate example. Conclusions are presented in Section 6.

2 Concept of a Multimodal Processes

The approach proposed is based on the LCP concept assuming its cyclic steady state behavior guaranteed by a given set of dispatching rules and assumed initial processes allocations. That means there exists a set of possible cyclic steady states encompassing potential cyclic behaviors of the LCP at hand. Each cyclic steady state specifies a local process periodicity, i.e. its cycle time. In that context, multimodal processes that can be seen as processes composed of local cyclic processes lead to the following fundamental questions that determine our further works:

- Does exist steady cyclic state of MCPs for the given LCPs structure constraints?
- Does exist a control procedure (i.e. a set of dispatching rules and an initial state) enabling to guarantee an assumed steady cyclic state (e.g. following requirements caused by MCP at hand) subject to LCPs structure constraints?

2.1 Illustrative Example

An idea of multiproduct production flow modeling, shown in Fig. 1, assumes a given layout of manufacturing system, i.e. machine tools and AGVS as well as structure of transportation path segments (see Fig. 1 b), and the LPCs model of local manufacturing routes following sequences of alternatively occurring machine tools and paths segments passed by AGVs (see Fig. 1 b).

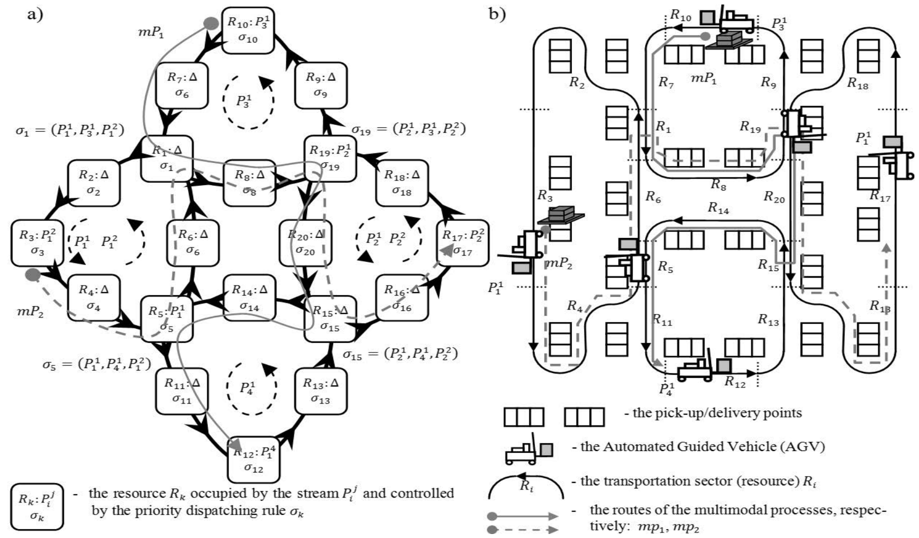


Fig. 1. Illustration of the LCP composed of four processes a), and modeling AGVS b)

In case considered the assumed cyclic (rotary) transportation routes can be seen as sequences:

$$p_1 = (R_1, R_2, R_3, R_4, R_5, R_6), \tag{1}$$

$$p_2 = (R_{15}, R_{16}, R_{17}, R_{18}, R_{19}, R_{20}), \tag{2}$$

$$p_3 = (R_1, R_8, R_{19}, R_9, R_{10}, R_7), \tag{3}$$

$$p_4 = (R_{11}, R_{12}, R_{13}, R_{15}, R_{14}, R_5), \tag{4}$$

The sequences (1), (2), (3), and (4) describe transportation routes of cyclic processes P_1, P_2, P_3, and P_4 executed in LCP network from Fig. 1 a). In general case, local processes can be seen as serial ones (i.e. more than one AGV executes its round trip service) processed along the same production route. In the case considered, processes P_1, P_2 are served by two AGVs each one, and specified by stream-processes P_1^1, P_1^2 and P_2^1, P_2^2, respectively (where P_i^j means the j-th stream of process P_i). That means, each the j-th stream-processes P_i^j (stream for short) modeling AGVs used in the i-th cyclic local transportation process P_i follows the same transportation route p_i.

In the LCP model from Fig. 1a), besides of local cyclic processes P_1, P_2, P_3, and P_4 one can assume the two new ones, e.g., mP_1, mP_2 described by:

$$mp_1 = (R_{10}, R_6, R_1, R_8, R_{19}, R_{20}, R_{15}, R_{14}, R_5, R_{11}, R_{12}), \tag{5}$$

$$mp_2 = (R_3, R_4, R_5, R_6, R_1, R_8, R_{19}, R_{20}, R_{15}, R_{16}, R_{17}). \tag{6}$$

These new routes distinguished in LCP model (see Fig. 1 a)) by the solid and dashed lines, respectively, provide the guidelines for two concurrently executed multimodal processes. Similarly to local processes, the multimodal processes that encompass pipeline flow of workpieces processed along the same production route can be also considered as cyclic and/or serial ones. In case from Fig. 1 a) processes mP_1, mP_2 consist of two streams each one, i.e. mP_1^1, mP_1^2 following the manufacturing route mp_1, and mP_2^1, mP_2^2 following the manufacturing route mp_2, respectively (where mP_i^j means the j-th stream of process mP_i. In order to simplify further considerations let us assume the AGVs servicing particular transportation routes are disposable for any machine tool in the manufacturing routes serviced.

2.2 Problem Formulation

Consider the digraph shown in Fig. 1 a). Four cycles specifying routes of local cyclic processes P_1, P_2, P_3 and P_4, respectively are distinguished. Each process route is specified by a sequence of resources passed on in course of its execution. The process routes are specified by (1), (2), (3), (4), where R_1, R_5, R_{15}, R_{19}, are so-called shared resources while the rest are non-shared ones. Processes sharing common resources interact each other on the base of mutual exclusion protocol.

The possible resources conflicts are resolved with help of priority rules determining the order in which streams of processes make their access to shared resources (for instance, in case of the resource R_5, the priority dispatching rule $\sigma_5 = (P_1^1, P_4^1, P_1^2)$ determines the order in which the stream P_1^1 can access the resource R_5 firstly, then P_4^1 as next and P_1^2, and once again P_1^1, and so on.

Consider two multimodal processes mP_1, mP_2 specified by the routes (5), (6). Since routes of multimodal processes consist of resources occurring in local process routes, hence multimodal processes can be seen as processes composed of sub-sequences of local processes. In other words each multimodal process can be treated as a process executing itself along distinguished parts of local processes determined by relevant routes sub-sections. In general case, each local process $P_i \in P = \{P_1, P_2, \ldots\ldots, P_n\}$, where: n – is a number of processes, the sequence of operations using resources defined by the given process route $p_i = \left(R_{j_1}, R_{j_2}, \ldots, R_{j_{lr(i)}}\right)$, $j_k \in \{1, 2, \ldots, m\}$, where: $lr(i)$ denotes a length of cyclic process route, m – is a number of resources, and $R_{j_k} \in R$, where $R = \{R_1, R_2, \ldots, R_m\}$, executes periodically. For each process P_i the set of streams Sr_i is assigned: $Sr_i = \left\{P_i^1, \ldots, P_i^{ls(i)}\right\}$. In that context, in a system from Fig. 1a) sets of streams are as follows: $Sr_1 = \left\{P_1^1, P_1^2\right\}$, $Sr_2 = \left\{P_2^1, P_2^2\right\}$,

$Sr_3 = \left\{P_3^1\right\}$, $Sr_4 = \left\{P_4^1\right\}$. In turn the set of all streams executed in the system is defined as: $SR = Sr_1 \cup \ldots \cup Sr_n$.

The sequence $T_i = \left(t_{i,j_1}, t_{i,j_2}, \ldots, t_{i,j_{lr(i)}}\right)$, $t_{i,j_k} \in N$, describes the operation times in each stream of P_i. To any shared resource $R_i \in R$ the priority dispatching rule $\sigma_i = \left(P_{j_1}, P_{j_2}, \ldots, P_{j_{lp(i)}}\right), P_{j_k} \in SR$ is assigned, where $lp(i) > 1, lp(i)$ – is a number of processes dispatched by σ_i. In general case the same streams can occur many times and in different orders so priority rule lengths are not limited. It means, in case of the following rule $\sigma_5 = \left(P_1^1, P_4^1, P_4^1, P_1^2\right)$ streams P_1^1, P_1^2 access uniquely while P_4^1 twice due to assumed order $P_1^1, P_4^1, P_4^1, P_1^2$.

Consider the set of multimodal processes $MP = mP_1, mP_2, \ldots, mP_u$, where u – is a number of multimodal processes. Similarly as in the case of local processes for each process mP_i the set of multimodal streams mSr_i is assigned: $Sr_i = \left\{mP_i^1, \ldots, mP_i^{mls(i)}\right\}$. In turn the set of all streams executed in the system is defined as: $MSR = mSr_1 \cup mSr_2 \cup \ldots \cup mSr_n$.

Each multimodal process mP_i is specified by the route mp_i which is a sequence of sub-sequences (sections) of local cyclic process routes:

$$mp_i = \left(mpr_j(a_j, b_j) \& \ldots \& mpr_h(a_h, b_h)\right), \tag{7}$$

where: $mpr_j(a, b) = (crd_a p_j, crd_{a+1} p_j, \ldots, crd_b p_j), crd_i D = d_i$,
for $D = (d_1, d_2, \ldots, d_i, \ldots, d_w), \forall a \in \{1, 2, \ldots, lr(i)\}, \forall j \in \{1, 2, \ldots, n\}, crd_a p_j \in R$. $u \& v = (u_1, \ldots, u_a, v_1, \ldots, v_b)$ denotes a concatenation of two sequences $u = (u_1, \ldots, u_a)$ and $v = (v_1, \ldots, v_b)$. Examples of multimodal routes are specified in (5), (6).

By analogy to local cyclic processes the sequence $mT_i = (mt_{i,j_1}, mt_{i,j_2}, \ldots,$ $\left.mt_{i,j_{lm(i)}}\right), mt_{i,j\,k} \in \mathbb{N}$ describes the operation times required by operations executed along mP_i (where $lm(i)$ length of i-th MCP route mP_i). In that context a LCP can be defined as a pair [2], [3]:

$$SC = (R, ST_{LCP}, SB_{LCP}), \tag{8}$$

where: $R = \{R_1, R_2, \ldots, R_m\}$ – the set of resources,

$ST_{LCP} = (P, SR, T, \mathit{\Pi}, \mathit{\Theta})$ – specifies the behavior of local processes, i.e.

$P = \{P_1, P_2, \ldots, P_n\}$ – the set of local process,

$SR = \left\{P_1^1, \ldots, P_1^{ls(i)}, P_2^1, \ldots, P_2^{ls(2)}, \ldots, P_n^1, \ldots, P_n^{ls(n)}\right\}$ – set of local processes streams,

$T = \{T_1, T_2, \ldots, T_n\}$ – the set of local processes operation times sequences,

$\mathit{\Pi} = \{p_1, p_2, \ldots, p_n\}$ – the set of local process routes, and $\mathit{\Theta} = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ – the set of dispatching priority rules.

$SB_{LCP} = (MP, MSR, M\mathit{\Pi}, MT)$ – characterizes the behavior of multimodal processes, i.e. $MP = \{mP_1, mP_2, \ldots, mP_u\}$ – the set of multimodal process, $MSR = \{mP_1^1, \ldots, mP_1^{mls(i)}, mP_2^1, \ldots, mP_2^{mls(2)}, \ldots, mP_n^1, \ldots, mP_n^{mls(n)}\}$ – set of multimodal processes streams, $M\mathit{\Pi} = \{mp_1, mp_2, \ldots, mp_u\}$ – the set of multimodal process routes, $MT = \{mT_1, mT_2, \ldots, mT_u\}$ – the set of multimodal process routes operations times.

Due to former assumption any operation time is equal to a unit operation time (1 u.t. for short), i.e.:

$$\forall_{i\in\{1,\ldots,n\}}\forall_{j\in\{1,\ldots,lr(i)\}}(crd_j T_i = 1), \tag{9}$$

$$\forall_{i\in\{1,\ldots,u\}}\forall_{j\in\{1,\ldots,lm(i)\}}(crd_j mT_i = 1). \tag{10}$$

In that context the problem considered can be stated as follows: Consider a LCP following the conditions (7). Given are initial allocations of both local and multimodal processes. The main question concerns of LCPs periodicity. In case the LCP behaves periodically the next question regards the LCPs period. Of course, the similar questions can be stated for multimodal processes executed in the LCP environment. Other questions regard of relationship between LCP and MCP periodicity, e.g. Does there exist the LCPs structure such that an assumed steady cyclic state the MCPs at hand can be achieved?

2.3 State Space

Consider the following MCPs state definition describing both the local and multimodal processes allocation:

$$mS^k = (S^r, MA^k), \tag{11}$$

where:

- $S^r \in \mathbb{S}l$ is the state of local processes, corresponding to the k-th state of multimodal processes,
$$S^r = (A^r, Z^r, Q^r), \tag{12}$$
where: $A^r = (a_1^r, a_2^r, \ldots, a_m^r)$ – the processes allocation in the r-th state, $a_i^r \in SR \cup \{\Delta\}$, $a_i = P_a^b$ – the i-th resource R_i is occupied by the b-th stream of process P_j, and $a_i^r = \Delta$ – the i-th resource R_i is unoccupied.
$Z^r = (z_1^r, z_2^r, \ldots, z_m^r)$ – the sequence of semaphores corresponding to the r-th state, $z_i^r \in SR$ – means the name of the stream (specified in the i-th dispatching rule σ_i, allocated to the i-th resource) allowed to occupy the i-th resource; for instance $z_i^r = P_a^b$ means that at the moment stream P_a^b is allowed to occupy the i-th resource.
$Q^r = ({q_1}^r, {q_2}^r, \ldots, {q_m}^r)$ – the sequence of semaphore indices, corresponding to the r-th state, ${q_i}^r$ determines the position of the semaphore ${z_i}^r$ in the priority dispatching rule σ_i, ${z_i}^r = crd_{({q_i}^r)}, \sigma_i$, ${q_i}^r \in \mathbb{N}$. For instance ${q_2}^r = 2$ and ${z_2}^r = P_1^1$, means the semaphore P_1^1 regards to position 2 in the priority dispatching rule σ_2.
- MA^k – the sequence of multimodal processes stream allocation: $MA^k = (mA_1^k, \ldots, mA_u^k)$, mA_i^k allocation of the process mP_i, i.e.:
$$mA_i^k = ({ma_{i,1}}^k, {ma_{i,2}}^k, \ldots, {ma_{i,m}}^k), \tag{13}$$
where: m – is a number of LCP resources, ${ma_{i,j}}^k \in mSr_i \cup \{\Delta\}, {ma_{i,j}}^k = mP_i^h \left(mP_i^h \in mSr_i\right)$ means, the j-th resource R_j is occupied by the h-th stream of multimodal process mP_i, and ${ma_{i,j}}^k = \Delta$ – the i-th resource R_j is unoccupied by streams of the i-th multimodal process P_i.

In that context, **the state $\mathbf{mS^k}$ is feasible** only if S^r is feasible [3] and for any of its ${ma_{i,j}}^k$ the following condition hold:

$$\forall_{i\in\{1,2,\dots,u\}}\exists!_{j\in\{1,2,\dots,m\}}\exists!_{h\in\{1,2,\dots,mls(i)\}}({ma_{i,j}}^k = mP_i^h), \tag{14}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{l|R_l\in\{r_a|r_a=crd_a mp_i, a=1,\dots,lm(i)\}\}}(ma_i, j^k \in mSr_i \cup \{\varDelta\}), \tag{15}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\notin\{l|R_l\in\{r_a|r_a=crd_a mp_i, a=1,\dots,lm(i)\}\}}(ma_i, j^k = \varDelta), \tag{16}$$

where: $lm(i)$ – the length of multimodal process route mP_i.

It means in every feasible state each multimodal process is allotted to a unique resource associated to the relevant multimodal process route. The introduced concept of the k-th state $\mathbf{mS^k}$ enables to create a space of feasible states $\mathbb{S}$. Moreover, the transition linking directly reachable states can be also introduced. Consider two feasible states mS^k and mS^l, i.e. $mS^k, mS^l \in \mathbb{S}$, such that:

$$mS^k = (S^{kr}, MA^k), \tag{17}$$

$$mS^l = (S^{kl}, MA^l). \tag{18}$$

The state $\mathbf{mS^l}$ **is reachable directly from the state** $\mathbf{mS^k}$ if the transition $S^{kr} \rightarrow S^{kl}$ holds, and the following conditions hold:

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{1,2,\dots,m\}}\left[({a_j}^{kr} = \varDelta) \Rightarrow ({ma_{i,j}}^k = {ma_{i,j}}^l)\right], \tag{19}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{1,2,\dots,m\}}\Big[[({a_j}^{kr} \neq \varDelta) \wedge ({a_j}^{kl} \neq \varDelta)] \Rightarrow ({ma_{i,j}}^k = {ma_{i,j}}^l)\Big], \tag{20}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{1,2,\dots,m\}}\Big[[({a_j}^{kr} \neq \varDelta) \wedge ({a_j}^{kl} = \varDelta) \wedge ({ma_{i,j}}^{kl} = \varDelta)] \Rightarrow ({ma_{i,j}}^l = \varDelta)\Big], \tag{21}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{1,2,\dots,m\}}\Big[[({a_j}^{kr} \neq \varDelta) \wedge ({a_j}^{kl} = \varDelta) \wedge ({ma_{i,j}}^{kl} \neq \varDelta) \wedge \Big(\alpha_j({a_j}^{kr}) = \alpha_j^*({ma_{i,j}}^k)\Big)] \Rightarrow [({ma_{i,j}}^l = \varDelta \wedge ({ma_{i,\alpha_j^*({ma_{i,j}}^k)}}^l = {ma_{i,j}}^k)]\Big], \tag{22}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{1,2,\dots,m\}}\Big[[({a_j}^{kr} \neq \varDelta) \wedge ({a_j}^{kl} = \varDelta) \wedge ({ma_{i,j}}^{kl} \neq \varDelta) \wedge \Big(\alpha_j^*({ma_{i,j}}^{kr}) = \varDelta\Big)] \Rightarrow [({ma_{i,j}}^l = \varDelta \wedge ({ma_{i,\lambda_j({ma_{i,j}}^k)}}^l = {ma_{i,j}}^k)]\Big], \tag{23}$$

$$\forall_{i\in\{1,2,\dots,u\}}\forall_{j\in\{1,2,\dots,m\}}\Big[[({a_j}^{kr} \neq \varDelta) \wedge ({a_j}^{kl} = \varDelta) \wedge ({ma_{i,j}}^{kl} \neq \varDelta) \wedge \Big(\alpha_j({a_j}^{kr}) \neq \alpha_j^*({ma_{i,j}}^k) \wedge (\alpha_j^*({ma_{i,j}}^k) \neq \varDelta)\Big)] \Rightarrow ({ma_{i,j}}^k = {ma_{i,j}}^l)\Big], \tag{24}$$

where: m – the number of resources, u – the number of multimodal processes, $\alpha_i(P_j^h)$ – an index of the resource directly succeeding the resource R_i, in the j-th local process route p_j, $\alpha_i(P_j^h) \in \{1, 2, \dots, m\}$, in case the resource R_i is the last element in the sequence p_j, then $\alpha_i(P_j^h)$ states for an index of the resource beginning these sequence,

$\alpha_j^*(mP_j^h)$ – an index of resource directly succeeding the resource R_i, in the j-th multimodal process route mp_j, $\alpha_j^*(mP_j^h) \in \{\mathit{\Delta}, 1, 2, \ldots, m\}$, in case the resource R_i is the last element in the sequence p_j, then $\alpha_j^*(mP_j^h) = \mathit{\Delta}$,

$\lambda_j(mP_j^h)$ – an index of the first resource in the j-th multimodal process route mp_j, $\alpha_j^*(mP_j^h) \in \{\mathit{\Delta}, 1, 2, \ldots, m\}$.

The conditions provided describe the relationship between subsequent allocations of multimodal processes MA^k and MA^l being in states mS^k and mS^l, respectively. Therefore, $mS^k \to mS^l$ linking two feasible states mS^k,$mS^l \in \mathbb{S}$ following the conditions (19)÷(24) can be seen as **a partial state transition function**, (describing the relationship between subsequent states in LCPs S^{kr} and S^{kl}.

Let us assume that multimodal process execution is conflict and deadlock-free as well that the executions of so called elementary subsequences depend on local processes, i.e. moments of their initiation depend on local processes "availability" (i.e. the moments the local and multimodal processes are allocated at the same resource).

3 Multimodal Processes Scheduling

The set $mSc^* = \{mS^{k_1}, mS^{k_2}, mS^{k_3}, \ldots, mS^{k_v}\}$, $mSc^* \subset \mathbb{S}$ is called **a reachability state space of multimodal processes** generated by an initial state $mS^{k_1} \in \mathbb{S}$. If the following conditions hold:

$$mS^{k_1} \xrightarrow{i-1} mS^{k_i} \xrightarrow{v-i-1} mS^{k_v} \to mS^{k_i} \tag{25}$$

where: $mS^a \xrightarrow{i} mS^b$ – the transition defined by (19) ÷ (24), as well as [3], (describing transitions between S^{kr} and S^{kl}).

The set $mSc = \{mS^{k_i}, mS^{k_{i+1}}, \ldots, mS^{k_v}\}$, $mSc \subseteq mSc^*$ is called **a cyclic steady state of multimodal processes** with the period $Tm = \|mSc\|$, $Tm > 1$. In other words a cyclic steady state contains such a set of states in which starting from any distinguished state it is possible to reach the rest of states and finally reach this distinguished state again:

$$\forall_{mS^k \in mSc} \left(mS^k \xrightarrow{Tm-1} mS^k \right) \tag{26}$$

Therefore, our former question regarding periodicity of MCP executed in LPCs results in the question whether there exist an initial state mS^0 generating the cyclic steady state mSc. It means, that searching for a cyclic steady state mSc in a given LPC can be seen as a reachability problem where for an assumed initial state mS^0 (i.e. determining local and multimodal processes allocations) the state mS^k, such that following transitions $mS^0 \xrightarrow{i} mS^k \xrightarrow{Tm-1} mS^k$ hold, is sought.

Note that for a given initial state mS^0, there exists a unique sequence of transitions leading to a cyclic steady state of both local and multimodal processes. So, starting from the state mS^0 the LCPs behavior results in a cyclic steady state. Moreover, it can be shown that steady state mSc exists if and only if the steady state Sc there exists. The following theorem provides the condition linking periods of LCP and MCP.

Theorem
Consider LCP and initial feasible state $mS^k = (S^{kr}, MA^k)$ such that $S^{kr} \in Sc$ and $mS^{kr} \in mSc$, where mSc – the cyclic steady state of multimodal processes, Sc the cyclic steady state of local processes, Tm – the period of mSc, Tl – the period of Sc. The following condition holds: $MOD(Tm, Tl) = 0$.

4 Declarative Modelling

Reachability problem can be stated in terms of Constraint Satisfaction Problem (CSP) [1], [9]. So, the mSc can be modeled in declarative framework as follows:

$$CS = ((X, D), C), \tag{27}$$

where: $X = \{Tm, mSc^*\}$ – the set of decision variables (the cycle of MCP, the set of states contained by the multimodal processes cyclic steady state), D the family of decision variable domains; C the set of constraints (28)÷(20); $Tm \in \mathbb{N}$, $mSc^* \subseteq \mathbb{S}$.

It means, the set mSc^*, see $X = \{Tm, mSc^*\}$, has to follow the constraints C:

$$Tm + i = \|mSc^*\|, i \geq 0, \tag{28}$$

$$\exists_{mS^k \in mSc^*}(mS^0 \xrightarrow{i} mS^k \xrightarrow{Tm-1} mS^k), \tag{29}$$

$$\exists!_{mSc \subseteq mSc^*} \forall_{mS^k \in mSc}(mS^k \xrightarrow{Tm-1} mS^k), \tag{30}$$

where: mS^0 – an initial state; $mS^a \xrightarrow{i} mS^b$ – the state transition following the conditions (19)÷(24).

Therefore, the problem (27) reduces to determination of the set X that follows constraints C, and can be implemented within the declarative languages environment. Note, the problem considered is formulated in a straight way. However, within the modeling framework employed the opposite one, i.e. reverse problem formulation can be considered.

In that case the variables $\{R, P, SR, MP, \Pi, M\Pi, MSR, T, MT, \Theta\}$ are treated as decision ones, and Tm, mSc^* as well as conditions (28)÷(30) state for constraints. For illustration let us consider the LCP and two MCP routes as shown in Fig. 2. The operation times of local and multimodal processes are equal to one (due to (9) and (10)). The response to the following questions is sought: Does the cyclic steady state of LCPs can be reached? What are Tl and Tm cycle times? What is tact time (cycle time) (i.e. the work time between two consecutive units completion) of the serial production case considered? What are the recommendations regarding the multi-load AGVs and the multi-load pick-up/delivery points (e.g. pallet carousel turntable like) usage? What are the changes imposed on the material handling transportation system (MHS)?

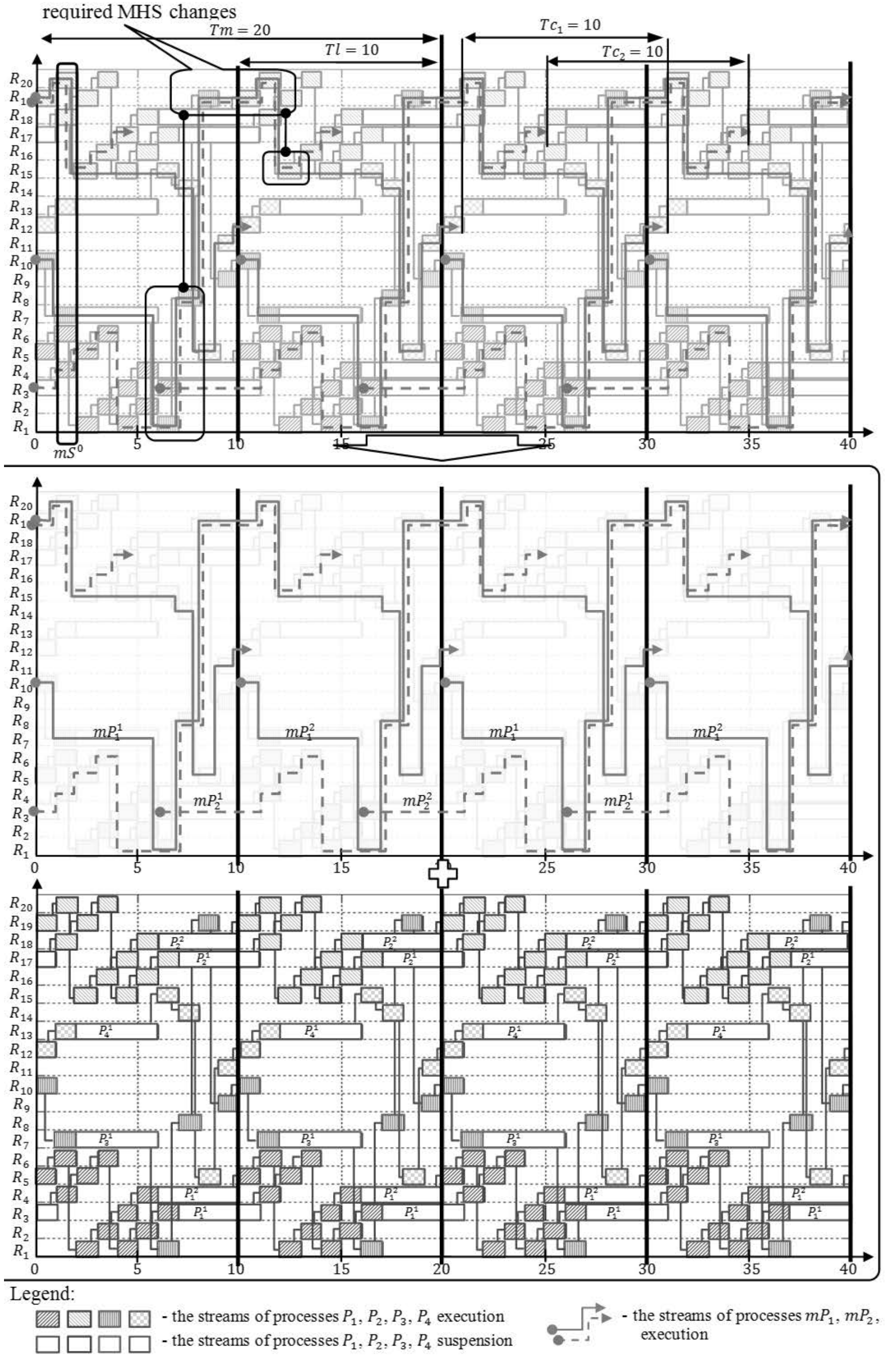


Fig. 2. Gantts charts illustrating execution of processes from. Fig. 1 a) , and modeling AGVS b)

Assuming the following initial state the response to these questions results from CS (27): $mS^0 = (S^0, MA^0)$, where:

- $S^0 = (A^0, Z^0, Q^0)$ – an initial state for local cyclic processes:
  $A^0 = (\Delta, \Delta, \Delta, P_1^1, \Delta, P_1^2, P_3^1, \Delta, \Delta, \Delta, \Delta, \Delta, P_4^1, \Delta, \Delta, \Delta, \Delta, P_2^1, \Delta, P_2^2)$,
  $Z^0 = (P_1^2, P_1^2, P_1^2, P_1^1, P_1^1, P_1^2, P_3^1, P_3^1, P_3^1, P_4^1, P_4^1, P_4^1, P_4^1, P_4^1, P_2^2, P_2^2, P_2^2, P_2^1, P_2^1, P_2^2)$,
  $Q^0 = (3, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 3, 2, 2, 1, 1, 1)$,
- $MA^0 = (mA_1^0, mA_2^0)$ – an initial allocation of multimodal processes:
  $mA_1^0 = (\Delta, \Delta, \Delta, \Delta, \Delta, \Delta, mP_1^1, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, mP_1^2)$,
  $mA_2^0 = (\Delta, \Delta, \Delta, mP_2^2, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta, mP_2^1)$.

The problem has been implemented in Oz Mozart. Its solution was found in ten steps, including both constraints propagation and variables distribution. The corresponding steady state period is: Tm=20 u.t. while the period of local processes is: Tl=10 u.t. That means the required mixture of two products on the system output can be awaited every 10 u.t. – tact times (time between streams execution) of processes mP_1, mP_2 equal to: $Tc_1 = Tc_2 = 10$ respectively. Moreover, the cyclic steady state mSc is equal to the multiplicity of cycle time generated by Sc (see the Theorem), see Fig. 2. Note that planned MCP schedule impose some changes in MHS structure, i.e. regards the double-load AGVs supporting P_2^2, P_3^1, and the double-load pick-up/delivery points modeled by $R_1, R_8, R_{15}, R_{19}, R_{20}$, see Fig. 1.

5 Concluding Remarks

In this paper we focused our attention to flow shops, job shops, assembly systems, and, in general, to any discrete event system which transforms raw material and/or components into products and/or components. Modeling and traffic control problem of a fleet of automated guided vehicles providing material handling/transportation service to the medium- and large-series manufacturing system which provides constantly the required mixture of various products on the system output, through the realizing so called cyclic manufacturing policy is its main contribution. The approach proposed is based on the system of concurrently flowing local cyclic processes LCPs concept. Its cyclic steady state behavior guaranteed by a given set of dispatching rules and assumed set of initial processes allocations provides the framework for MCPs modeling and scheduling. The MCP concept lead to two fundamental questions: Does there exist a control procedure (i.e. a set of dispatching rules and an initial state) enabling to guarantee an assumed steady cyclic state (e.g. following requirements caused by MCPs at hand) subject to LCPs structure constraints? Does there exist the LCPs structure such that an assumed steady cyclic state (e.g. following requirements caused by MCPs at hand) can be achieved? Response to these questions determines our further works.

We believe that this approach enables us to develop the sufficient conditions that allow one to compose elementary systems in such a way as to obtain the final MCPs scheduling system with required quantitative and qualitative characteristics. So, we are looking for a method allowing one to replace the exhaustive search for the admissible control by a step-by-step structural design guaranteeing the required system behavior.

References

1. Bocewicz, G., Wójcik, R., Banaszak, Z.A.: Cyclic steady state refinement. In: Abraham, A., Corchado, J.M., González, S.R., De Paz Santana, J.F. (eds.) International Symposium on DCAI. AISC, vol. 91, pp. 191–198. Springer, Heidelberg (2011)
2. Bocewicz, G., Wójcik, R., Banaszak, Z.: Design of admissible schedules for AGV systems with constraints: A logic-algebraic approach. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 578–587. Springer, Heidelberg (2007)
3. Bocewicz, G., Banaszak, Z.: Declarative approach to cyclic steady states space refinement: periodic processes scheduling. The International Journal of Advanced Manufacturing Technology 67(1-4), 137–155 (2013)
4. Fournier, O., Lopez, P., Lan Sun Luk, J.-D.: Cyclic scheduling following the social behavior of ant colonies. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp. 450–454 (2002)
5. Kats, V., Lei, L., Levner, E.: Minimizing the cycle time of multiple-product processing networks with a fixed operation sequence, setups, and time-window constraints. European Journal of Operational Research 187, 1196–1211 (2008)
6. Liebchen, C., Möhring, R.H.: A case study in periodic timetabling. Electronic Notes in Theoretical Computer Science 66(6), 21–34 (2002)
7. Levner, E., Kats, V., Alcaide, D., Pablo, L., Cheng, T.C.E.: Complexity of cyclic scheduling problems: A state of-the-art-survey. Computers & Industrial Engineering 59(2), 352–361 (2010)
8. Polak, M., Majdzik, P., Banaszak, Z., Wójcik, R.: The performance evaluation tool for automated prototyping of concurrent cyclic processes. Fundamenta Informaticae 60(1-4), 269–289 (2004)
9. Sitek, P., Wikarek, J.: A Declarative Framework for Constrained Search Problems. In: Nguyen, N.T., Borzemski, L., Grzech, A., Ali, M. (eds.) IEA/AIE 2008. LNCS (LNAI), vol. 5027, pp. 728–737. Springer, Heidelberg (2008)
10. Słowik, A.: Steering of Balance between Exploration and Exploitation Properties of Evolutionary Algorithms - Mix Selection. In: Rutkowski, L., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2010, Part II. LNCS (LNAI), vol. 6114, pp. 213–220. Springer, Heidelberg (2010)
11. Smutnicki, C.: Cyclic job shop scheduling, Technical Report PRE 44/2009, Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wrocław (2009)
12. Song, J.-S., Lee, T.-E.: Petri net modeling and scheduling for cyclic job shops with blocking. Computers & Industrial Engineering 34(2), 281–295 (1998)
13. Trouillet, B., Korbaa, O., Gentina, J.-C.K.: Formal Approach for FMS Cyclic Scheduling. IEEE SMC Transactions, Part C 37(1), 126–137 (2007)
14. Wang, B., Yang, H., Zhang, Z.-H.: Research on the train operation plan of the Beijing-Tianjin intercity railway based on periodic train diagrams. Tiedao Xuebao/Journal of the China Railway Society 29(2), 8–13 (2007)
15. Von Kampmeyer, T.: Cyclic scheduling problems. Ph.D. Dissertation, Fachbereich Mathematik/Informatik, Universität Osnabrück (2006)

Integrated Scheduling of Quay Cranes and Automated Lifting Vehicles in Automated Container Terminal with Unlimited Buffer Space

Seyed Hamidreza Sadeghian[1,2,*], Mohd Khairol Anuar bin Mohd Ariffin[1],
Tang Sai Hong[1], and Napsiah bt Ismail[1]

[1] Department of Mechanical and Manufacturing Engineering, Engineering Faculty,
Universiti Putra Malaysia, Malaysia
{khairol,saihong,napsiah}@eng.upm.edu.my
[2] Department of Industrial Engineering, Islamic Azad University-Lenjan Branch,
Isfahan, Iran
sadeghian@iauln.ac.ir

Abstract. Nowadays, role of sea port container terminals in national and regional transportation and economy cannot be omitted. To respond enormous and every increasing demand on sea transshipments within the same time frame, terminal managers require more and more efficiency in container performance and operations. Automation of the processes at the quays of the container ports is one the solutions to improve the performance and output of container terminals. For such purpose, using new generation of vehicles is unavoidable. Automated Lifting Vehicle (ALV) is one of the automatic vehicles that has been introduced during recent years and can be used in container terminals. In this paper, an integrated scheduling of quay cranes and automated lifting vehicles with unlimited buffer space is formulated as a mixed integer linear programming model. Our objective is to minimize the makespan of all the loading and unloading tasks for a pre-defined set of cranes. Obtained result from our scheduling model is compared with an Automated Guided Vehicle (AGV) inspired from the same problem.

Keywords: Automated container terminal, Quay crane, Automated lifting vehicle, Unlimited buffer space, Integrated scheduling.

1 Introduction

Maritime transport is one the essential supports for globalization and a huge portion of international trade is being transported through the ports. Due to the significant role of maritime transports major ports are expected to increase their cargo capacity to two or three times more by 2020 [1].

Containers are suitable, safe, secure and efficient carriers for storage and shipping of products and materials in sea transport. A shipping container is a box that is designed for door to door delivery of the goods without physical handling of the

[*] Corresponding author.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_58, © Springer International Publishing Switzerland 2014

contents [2]. Container as a necessary part of a unit load concept has achieved a certain place in global sea cargo transportation. Nowadays, containers transport more than 60% of the world's deep-sea general cargo, especially between economically stable and strong countries [3]. As a result of the continuing increase of container trade, many sea terminals are equipped to serve the containerships and competition between major seaports and container terminals is becoming more and more. So it is important for port operators to develop different optimization algorithms and decision tools to improve their performance and competitiveness. The competitiveness of a container seaport is defined by different success factors, especially fastness of the loading and unloading activities. So it is essential that a terminal can receive, store, and dispatch containers efficiently and rapidly [4].

To increase efficiency of the container terminals, it is necessary to coordinate different terminal equipment to ensure a correct flow of containers within the terminal. Container activities can be categorized into: export, import and transshipment activities. In export activities, the containers are being shipped and stored at their predefined locations in the storage yard. For loading the containers, yard cranes (YC) will retrieve them from the stored locations and vehicles transport the containers to the quay side. Then quay cranes (QC) receive containers from the vehicles and load them into the vessels. The processes for import activities are performed in the same manner but in the reverse order. For transshipment activities, after unloading from the vessel, containers will be stored in the storage yard and finally be loaded onto other vessels.

Problems related to operations and activities in container terminals can be divided into several types of problems, such as assignment of vessels [5], loading or discharging and storage of the containers in marshaling yard [5], scheduling of quay cranes [6], planning of YCs [7] and assignment of storage places to containers [8].

In automatic container terminals, several types of vehicles can be used for handling and transferring the containers in the yard. Two different types of automatic vehicles that being used are: Automated Guided Vehicle (AGV) and Automated Lifting Vehicle (ALV). An AGV can receive a container from quay crane and transport it over a fixed path. In such situation, a yard crane should take the container off the vehicle. ALVs are capable of lifting a container from the ground by themselves. Because of such capability, in terminal, buffer areas are defined at QC in apron and transfer point (TP) in the yard to help loading and unloading process for ALVs. ALV receives the container from a buffer area and carries it to its destination.

Compared to AGVs, only few prior researches have involved ALVs. Vis et al. [9] has compared the performance of AGV and ALV, as two types of known automated vehicles, by a simulation study. They concluded that, by observing purchasing costs and initial essential investment for equipments, ALVs are the cheaper options than AGVs (in some cases 38% less ALVs need to be used than AGVs). Nguyen and Kim, [10] developed a mixed programming model for the optimal assignment of delivery tasks to ALVs. They have proposed a heuristic algorithm to solve their model. Le et al. [11] have used DCA for solving their model.

In this work, the authors consider some of constraints similar to the Nguyen's model [10]. Minimizing makespan is objective of our model that is also used by Homayouni et al. [12] for dispatching of AGVs in container's terminal. In the proposed mixed zero-one programming model of this study, unlimited buffer space for QCs is considered and so the delay of ALVs and QCs for lack of empty buffer space will not occur.

2 Problem Definition and Mathematical Model

2.1 Problem Definition

The handling activities can be divided into two parts, one portion of these activities which are performed by QCs are known as seaside operations, and another part that will be done by ALVs is called landside operations.

Before starting of ship operations, shipping agent prepares a guideline for loading and unloading operations based on the schedule of QCs. According to the guideline and work schedule, a sequence list will be issued that determines the sequence of unloading and loading operations for all the containers. In most of the times, actual ship operations follow the specified order in sequence list. So we can consider that sequence and delivery operations of ALVs are predefined and known in advance.

The function and duty of an ALV for unloading tasks is delivering a container from the apron to the yard, and for loading operations it should carry the container from the yard to the apron. During the unloading operation, QC picks up a container from the prow and delivers it into the buffer space. In container terminals with limited buffer spaces, when the buffer is full, ALV or QC must wait for releasing a container on buffer space. In our problem, we have considered unlimited buffer space for QCs and therefore delay of ALVs and QCs for lack of empty buffer space is eliminated. When QC delivered the container to the buffer, ALV picks up the container from the buffer and delivers it to the marshalling yard. In the marshalling yard, ALV releases the container to the specified and available transfer point (TP) of the yard. An AYC picks up and stacks container onto an empty and predefined place in bay. The loading operation is performed in the reverse order.

2.2 Mathematical Model

During developing the model, the authors assumed that YCs are not known as the bottle neck of the container terminal. It means that the vehicles can be served by the YCs immediately, and yard cranes are ready to pick up the imported containers without any delay. Also the exported containers are ready and available to be delivered to the ALV while it reaches the loading or unloading place.

The ALV's journey starts from predetermined loading/unloading station and finishes with coming back to the initial position. In the proposed model we have assumed that QCs are far enough from each other and there is enough and unlimited space for buffers in apron. In other words, quay cranes and ALVs can release the container to buffer as soon as reach the place. Some other assumptions in the formulation of the problem are as follows:

- Each ALV transports only one container at each time.

- All ALVs are same in capacity and shape, thus they are neither assigned to a specific kind of container nor to a crane.
- ALV's Congestions are not considered in the model.
- Operation time of ALV or QC for pick up and releasing the container is small enough and can be neglected.
- Travel times of ALVs, travel time of cranes between the quay and the vessel area (TQ) and its operation time (OQ) is deterministic and predefined.

The following notations are used in the proposed Mixed-Integer Programming (MIP) model for dispatching of ALVs:

V The set of ALVs.

K The set of QCs.

m_k The number of tasks determined for QC_k , $k \in K$.

m_l The number of tasks for ALV_l , $l \in V$.

T_i^k The ith operation of QC_k , $k \in K$, $i = 1, \dots, m_k$.

y_i^k The real completion time of T_i^k, $k \in K$, $i = 1, \dots, m_k$.

s_i^k The earliest possible completion time of T_i^k, $k \in K$, $i = 1, \dots, m_k$.

K $\{0\} \cup K$.

$K^{"}$ $\{F\} \cup K$.

C_j Cycle time of ALV_j , $j \in V$.

c_{ki}^{lj} The travel time between QC_k and QC_l including required time for the ALV to be ready for T_j^l after it experiences T_i^k, $k \in K', l \in K^{"}$, $i = 1,2, \dots, m_k$, $j = 1,2, \dots, m_l$.

x_{ki}^{lj} The decision variable that becomes 1 if T_j^l be executed directly after T_i^k by the same ALV, $k \in K'$, $l \in K^{"}$, $i = 1,2, \dots, m_k$, $j = 1,2, \dots, m_l$.

M A big positive number.

OQ The operational time of quay cranes.

TQ The travel time of quay cranes between the ship and the quay area.

L The set of loading tasks.

U The set of Unloading tasks.

The problem of scheduling of lifting vehicle to transfer containers in an automated port container terminal with unlimited buffer space is a static scheduling and assignment problem for ALVs to accomplish all the delivery tasks without any limitation for buffer capacity. The objective function of the developed model is as below:

$$\text{Minimize } Z = Makespan \tag{1}$$

Different objective functions can be defined to improve the transfer and traveling of containers but in this model we have focused on minimizing the makespan of all loading and unloading tasks in a specific scheduling horizon (1). The makespan of tasks is the completion time for latest journey of the ALVs to the final destinations. Minimizing the makespan of ALVs will result in decreasing the completion time and delay of the quay cranes. Constraints for this model are described as follows:

$$\begin{gathered} \mathrm{C_j} - \left(\mathrm{y_i^k} - \mathrm{TQ} - \mathrm{OQ} + \mathrm{c_{ki}^{fj}}\right) \geq \mathrm{M}\left(\mathrm{x_{ki}^{fj}} - 1\right), \forall \mathrm{k} \in \mathrm{K}, \mathrm{j} \in \mathrm{V}, \\ \mathrm{i} = 1,2,\dots,\mathrm{m_k}, \ \mathrm{T_i^k} \in \mathrm{L} \end{gathered} \tag{2}$$

$$\mathrm{C_j} - \left(\mathrm{y_i^k} + \mathrm{c_{ki}^{fj}}\right) \geq \mathrm{M}\left(\mathrm{x_{ki}^{fj}} - 1\right), \forall \mathrm{k} \in \mathrm{K'}, \mathrm{j} \in \mathrm{V}, \mathrm{i} = 1,2,\dots,\mathrm{m_k}, \ \mathrm{T_i^k} \in \mathrm{U} \tag{3}$$

$$\text{Makespan} \geq \mathrm{C_j} \qquad \forall \mathrm{j} \in \mathrm{V} \tag{4}$$

$$\sum_{\mathrm{l} \in \mathrm{K''}} \sum_{\mathrm{j}=1}^{\mathrm{m_l}} \mathrm{x_{ki}^{lj}} = 1, \forall \mathrm{k} \in \mathrm{K'}, \mathrm{i} = 1,2,\dots,\mathrm{m_k} \tag{5}$$

$$\sum_{\mathrm{k} \in \mathrm{K'}} \sum_{\mathrm{i}=1}^{\mathrm{m_k}} \mathrm{x_{ki}^{lj}} = 1, \forall \mathrm{k} \in \mathrm{K''}, \mathrm{j} = 1,2,\dots,\mathrm{m_l} \tag{6}$$

$$\mathrm{y_i^k} \geq \mathrm{s_i^k}, \forall \mathrm{k} \in \mathrm{K'}, \mathrm{i} = 1,2,\dots,\mathrm{m_k} \tag{7}$$

$$\mathrm{y_{i+1}^k} - \mathrm{y_i^k} \geq \mathrm{s_{i+1}^k} - \mathrm{s_i^k}, \forall \mathrm{k} \in \mathrm{K}, \mathrm{i} = 1,2,\dots,\mathrm{m_{k-1}} \tag{8}$$

$$\begin{gathered} \mathrm{y_j^l} - \mathrm{A} \geq \mathrm{M}\left(\mathrm{x_{ki}^{lj}} - 1\right), \forall \mathrm{k} \in \mathrm{K'}, \forall \mathrm{l} \in \mathrm{K''}, \forall \mathrm{i} = 1,2,\dots,\mathrm{m_k}, \\ \forall \mathrm{j} = 1,2,\dots,\mathrm{m_l} \end{gathered} \tag{9}$$

$$\mathrm{x_{ki}^{lj}} = 0 \text{ or } 1, \forall \mathrm{k} \in \mathrm{K'}, \mathrm{l} \in \mathrm{K''}, \mathrm{i} = 1,2,\dots,\mathrm{m_k}, \mathrm{j} = 1,2,\dots,\mathrm{m_l} \tag{10}$$

Constraints (2) and (3) define the cycle time of ALVs, including the time that the ALV delivers the last container to destination, and the travel time of its last journey to the assigned final location. In loading operations, the quay crane continues its task after receiving the container and the ALV is allowed to continue its travel, so for calculation of cycle time, *OQ* and *TQ* should be deducted. Depend on the current duty and previous assigned task to a specific ALV, c_{ki}^{lj} can be different. More details for calculation of c_{ki}^{lj} are presented in Table 1. In this table *S, F, L* and *U* represent the

Start, Finish, Loading and Unloading tasks and *a*, *b* and *c* are ALV traveling times between QCs and TPs.

Table 1. Calculation for c_{ki}^{lj}

T_i^k	T_j^l	c_{ki}^{lj}
S	L	$a+b$
S	U	a
L	L	$a+b$
L	U	a
L	F	a
U	L	$a+b+c$
U	U	$a+b$
U	F	$a+b$

Constraint (4) shows that makespan is the largest cycle time of the ALVs calculated through formula (2) and (3). Constraints (5) and (6) ensure a one to one relation between two sequential tasks including the initial and final journeys of the ALVs. Constraint (7) expresses that the actual completion time is always greater than or equal to the earliest possible completion time. Constraint (8) defines that between two tasks assigned to a specific QC, there should be enough time for the QC to perform all the required movements. Constraint (9) indicates that the y_j^l is depended on y_{j-1}^l and y_i^k. In other words, completion time of T_j^l on QC_j is related to previous duty of the ALV and completion time of prior assigned task to the QC. Based on current operation of QC and different characteristics of the T_i^k and T_{j-1}^l , this parameter varies. More detailed calculation for y_j^l is presented in Table 2. The "Max" function in this constraint can be separated into two inequalities to make a linear set of constraints. Constraint (10) defines x_{ki}^{lj} as binary decision variable.

In this model *Makespan,* C_j and y_i^k will be obtained during solving the model and through the calculations depend on which x_{ki}^{lj} s get 1 value and which one be 0. A feasible solution is a one to one assignment between all the start and finish sets, represented by a series of x_{ki}^{lj} s. The start set is included starting events of ALVs and events related to the transfer operations by ALVs. And the finish set includes the stopping events of ALVs and events for delivery tasks of ALVs.

Table 2. Calculation for Constraint (9) on *A*

T_i^k	T_{j-1}^l	T_j^l	A
L	L	L	$Max(y_{j-1}^l + TQ, y_i^k - TQ - OQ + c_{ki}^{lj}) + TQ + OQ$
U	L	L	$Max(y_{j-1}^l + TQ, y_i^k + c_{ki}^{lj}) + TQ + OQ$
L	U	L	$Max(y_{j-1}^l, y_i^k - TQ - OQ + c_{ki}^{lj}) + TQ + OQ$
U	U	L	$Max(y_{j-1}^l, y_i^k + c_{ki}^{lj}) + TQ + OQ$
U/L	L	U	$y_{j-1}^l + TQ + OQ$
U/L	U	U	$y_{j-1}^l + 2TQ + OQ$

3 Numerical Experiments and Discussion

For comparison of the proposed model for dispatching of ALVs with unlimited buffer space by the same problem with AGV, a set of test cases is considered. 10 test cases are planned in a typical automated container terminal containing six transfer points in yard and six quay cranes in apron. In the generated test cases, the number of operations for each QC, the number of QCs and the number of ALVs range from 4 to 7, from 2 to 3 and from 3 to 4, respectively.

The travel times of ALVs between all combinations of QCs and TPs shown in Table 3 are same as traveling times that presented by Lau and Zhao [13].

Table 3. ALV traveling times between combinations of QCs and TPs (s) [13]

	QCs						TPs					
	0	1	2	3	4	5	6	7	8	9	10	11
0	022	302	602	902	120	150	155	852	115	145	175	205
1	80	0	30	60	90	120	85	55	85	115	145	175
2	110	80	0	30	60	90	115	85	55	85	115	145
3	140	110	80	0	30	60	145	115	85	55	85	115
4	170	140	110	80	0	30	175	145	115	85	55	85
5	200	170	130	110	80	0	205	175	145	115	85	55
6	55	85	115	145	175	205	0	80	110	130	170	200
7	85	115	145	175	205	235	30	0	80	110	140	170
8	115	145	175	205	235	265	60	30	0	80	110	140
9	145	175	205	235	265	295	90	60	30	0	80	110
10	175	205	235	265	295	325	120	90	60	30	0	80
11	205	235	265	295	325	355	150	120	90	60	30	0

The *OQ* for unloading and loading tasks is set to 20 s and the *TQ* is equal to 10 s for loaded or empty journeys. Table 4 shows details of test cases and the comparative results for ALV and AGV. In the first column, number of tasks, number of QCs, number of TPs and number of ALVs for each case are presented. Defined sequence of loading and unloading tasks for each quay crane and objective value for ALV and AGV are shown in column 3, 4 and 5. Also, the objective values are compared in the last column. All the tests for the ALV and AGV were solve by branch and bound algorithm and programmed in Lingo® software.

Table 4. Test cases and comparative results

T-QC-TP-ALV	QC No.	Task Type	ALV (A)	AGV (B)	Ratio (=A/B)
8-2-2-3	1,2	U,U,U,L; U,L,U,U	150	315	0.47619
8-2-2-4	1,2	U,U,U,L; U,L,U,U	125	290	0.43103
10-2-2-3	4,5	U,U,L,L,L; L,U,L,U,L	190	245	0.77551
10-2-2-4	4,5	U,U,L,L,L; L,U,L,U,L	150	205	0.73171
12-2-2-3	2,3	U,U,L,L,U,L;U,L,U,L,U,L	290	435	0.66667
12-2-2-4	2,3	U,U,L,L,U,L;U,L,U,L,U,L	220	365	0.60274
12-3-2-3	2,3,4	U,U,L,L;U,L,U,L;U,L,U,L	160	395	0.40506
12-3-2-4	2,3,4	U,U,L,L;U,L,U,L;U,L,U,L	130	330	0.39394
14-2-2-3	2,3	U,U,L,L,U,L,U;L,U,L,U,L,L,U	250	360	0.69444
14-2-2-4	2,3	U,U,L,L,U,L,U;L,U,L,U,L,L,U	195	320	0.60937

From numerical results, and as it can be seen in Fig.1 we observe that in all the test cases, ALV with unlimited buffer space has better results than AGV and in each case, as we expected, by increasing number of ALVs we have better and less makespan.

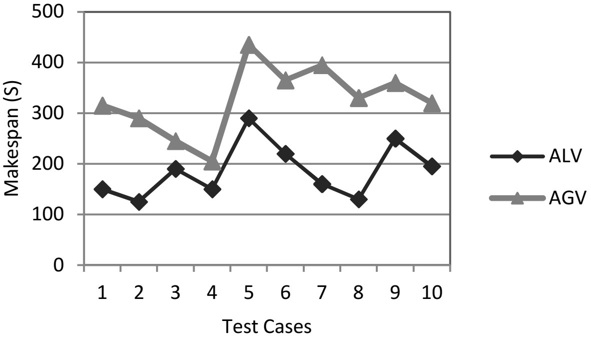


Fig. 1. Comparative results of test cases

4 Conclusion

This paper developed and discussed a static model for dispatching of ALVs to load or unload a predetermined number of containers in automated terminals with unlimited buffer spaces. The problem was formulated as a Mixed Integer Linear Programming (MILP) model to minimize the makespan of all transport tasks. The makespan is largest cycle time among the all ALVs to perform their assigned journeys from the initial locations to the final destinations. This objective function will decrease both the completion time of the QC tasks and ALV's traveling time. The authors considered test cases to evaluate performance of their model and compare the results by same problems with AGV. The obtained result shows that in all considered cases, ALV with unlimited buffer spaces has better performance than AGV.

References

1. Liu, C.I., Ioannou, P.A.: A comparison of different dispatching rules in an automated container terminal. In: IEEE 5th International Conference on Intelligent Transportation System, Singapore, pp. 880–885 (2002)
2. Cheng, Y.L., Sen, H.C., Natarajan, K., Teo, C.P., Tan, K.C.: Dispatching automated guided vehicles in a container terminal. Supply Chain Optimization, 355–389 (2005)
3. Steenken, D., VoB, S., Stahlbock, R.: Container terminal operation and operations research- a classification and literature review. OR Spectrum 26(1), 3–49 (2004)
4. Das, S.K., Spasovic, A.: Scheduling material handling vehicles in a container terminal. Production Planning & Control 14(7), 623–633 (2003)
5. Imai, A., Sasaki, K., Nishimura, E., Papadimitriou, S.: Multi-objective simultaneous stowage and load planning for a container ship with container rehandle in yard stacks. Er. J. Oper. Res. 171(2), 373–389 (2006)
6. Lee, D.H., Cao, J.X., Shi, Q.X.: Integrated quay crane and yard truck schedule for inbound containers. In: IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2008, Singapore, pp. 1219–1223 (2008)
7. Lee, D.H., Cao, Z., Meng, Q.: Scheduling of two-transtainer systems for loading outbound containers in port container terminals with simulated annealing algorithm. Int. J. Product. Econom. 107(1), 115–124 (2007)
8. Lee, L.H., Chew, E.P., Tan, K.C., Han, Y.: An optimization model for storage yard management in transshipment hubs. OR Spectrum 28(4), 539–556 (2006)
9. Vis, I.F.A., Harika, I.: Comparison of vehicle types at an automated container terminal. OR Spectrum 26(1), 117–143 (2004)
10. Nguyen, V.D., Kim, K.H.: A dispatching method for automated lifting vehicles in automated port container terminals. Comput. Ind. Eng. 56(3), 1002–1020 (2009)
11. Le, H., Yassine, A., Moussi, R.: DCA for solving the scheduling of lifting vehicle in an automated port container terminal. Computational Management Science 9(2), 273–286 (2012)
12. Homayouni, S.M., Tang, S.H., Ismail, N., Ariffin, M.K.A.: Using simulated annealing algorithm for optimization of quay cranes and automated guided vehicles scheduling. International Journal of Physical Sciences 6(27), 6286–6294 (2011)
13. Lau, H.Y.K., Zhao, Y.: Integrated scheduling of handling equipment at automated container terminals. International Journal of Production Economics 112(2), 665–682 (2008)

Genetic Algorithm Solving the Orienteering Problem with Time Windows

Joanna Karbowska-Chilinska and Pawel Zabielski

Bialystok University of Technology, Faculty of Computer Science, Poland

Abstract. The Orienteering Problem with Time Windows (OPTW) is a well-known routing problem in which a given positive profit and time interval are associated with each location. The solution to the OPTW finds a route comprising a subset of the locations, with a fixed limit on length or travel time, that maximises the cumulative score of the locations visited in the predefined time intervals. This paper proposes a new genetic algorithm (GA) for solving the OPTW. We use specific mutation based on the idea of insertion and shake steps taken from the well-known iterated local search method (ILS). Computational experiments are conducted on popular benchmark instances. The tests show that repetition of the mutation step for the same route during one iteration of GA can improve the solution so that it outperforms the ILS result.

Keywords: routing problem, orienteering problem with time windows, genetic algorithm.

1 Introduction

The simpler version of the Orienteering Problem with Time Windows (OPTW) is the Orienteering Problem (OP) [5]. In the OP [14], [9], a profit is associated with each location and a travel length or time is assigned to each linked pair of locations. The goal is to find a route with a fixed limit on length or travel time, containing a subset of the locations, which maximises the profit of the locations visited. In the OPTW a time window $[O_i, C_i]$ is additionally assigned to each location, where O_i and C_i denote the opening and closing time, respectively. Service of a location must begin and end within this interval. It is permissible to wait before the opening time in order to visit profitable locations and maximise the total score. A feasible solution will not violate any time window constraint on the fixed travel time limit (or travel length) of the route.

Numerous examples of practical applications of the OPTW are described in the literature, e.g. in logistics and production scheduling [13]. OPTW solutions are useful in solving problems related to tourism such as the Tourist Trip Design Problem (TTDP) [15], which involves generating the optimum route that complies with given constraints. Electronic tourist guides have enjoyed great popularity, particularly in the last few years [3], [15]. These devices select points of interests (POI) that maximise tourist satisfaction by taking into account tourist preferences (travel cost or distance, duration or starting time of the trip) as

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_59, © Springer International Publishing Switzerland 2014

well as the opening and closing times of the POI. Other important extensions of the OPTW having application in the TTDP include the Team Orienteering Problem with Time Windows (TOPTW) and the Time Dependant Orienteering Problem with Time Windows (TDOPTW). The TOPTW [1] expands the OPTW to include multiple tours, with each tour satisfying the same fixed travel length or time constraint. In the TDOPTW the travel time between locations is not fixed but varies in time, i.e. the travel time from location i to j depends on the departure time from i. Thus the TDOPTW could serve as an excellent model for a TTDP problem in which tourists use public transport [3].

Researchers have investigated two categories of OPTW solutions: exact solutions and heuristic approaches. Righini and Salani [11] developed an exact algorithm based on bi-directional dynamic programming. The OPTW (like the OP) is an NP-hard problem [5], so various heuristic approaches are usually used in practical applications, e.g. a granular variable neighbourhood search (GVNS) [4] or the ant colony optimisation approach (ACO) [10]. The iterated local search (ILS) [16] is the fastest known heuristic and is used to solve the Tourist Trip Design Problem [3]. The heuristic iteratively builds one route combining an insertion step and deletion of some consecutive locations (shake step) to escape from a local optimum.

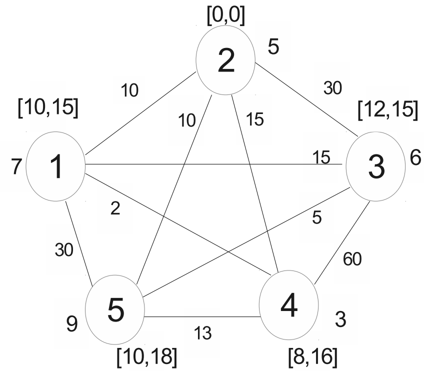
In papers [6], [8], [7] an effective genetic algorithms solving the OP problem was introduced. In this paper we introduce time windows to the genetic algorithm (GA). The GA is based on the method we proposed in [7]. Here we propose a novel method for generating an initial solution. Furthermore, we use specific crossover and mutation, which is based on the idea of insertion and shake steps taken from the ILS method. The main contribution of this paper is an algorithm that uses repetition of the mutation step for the same route during one iteration and obtains better score results than the ILS method. By score result we mean the sum of the profits assigned to the locations on the route generated.

The remainder of the paper is organised as follows. The problem definition and an example are presented in Section 2. In Section 3, we describe the concept of our genetic algorithm for solving the OPTW. The results of computational experiments run on benchmark datasets are discussed in Section 4. In Section 5 we present our conclusions and lay the groundwork for further research.

2 The Problem Definition

The Orienteering Problem with Time Windows (OPTW) can be described as a graph optimisation problem as follows. Let G be a graph with n vertices, where each vertex has a profit p_i, a visiting time T_i and a time window $[O_i, C_i]$, where O_i and C_i denote the opening and closing time, respectively. Each edge between vertices i and j has a fixed cost t_{ij} (interpreted as the time or length needed to travel between locations). The starting point s and ending point e are given as well. The OPTW goal is to determine a single route that visits some of vertices within the fixed time windows and maximises the total profit. Moreover, the total cost of the edges on the path must be less than the threshold t_{max}, and any vertex on the path can only be visited once. It is permissible to wait at a vertex before its time windows opens.

A simple example of a graph with five vertices is shown in Figure 1. The travel time values (given in minutes) are indicated on the edges. A profit value and a time window $[O_i, C_i]$ are marked next to each vertex. In the table next to the graph a visiting time (in minutes) is given for each vertex. Let vertex 1 be the starting point of the route at 10 am and vertex 5 the ending point at 5 pm. The value of t_{max} is 7 hours. The solution to the OPTW in the example is the route 1-2-3-5 with the total profit 27 and travel time equal to 60+10+120+30+60+5+60=345 minutes = 5 hours 45 minutes.



i	visiting time *Ti*
1	60
2	120
3	60
4	90
5	60

Fig. 1. Graph example to illustrate the OPTW problem

3 Genetic Algorithm

To solve the OPTW, the following genetic algorithm (GA) is used. First, a population of P_{size} individuals (routes) is generated from the starting point to the ending point. The value of the fitness function F is calculated for each route to estimate their quality. In our GA we use F as in [6], [7], which is equal to $TotalProfit^3/TravelTime$. It takes into account the sum of the profits assigned to the vertices on the route and the total travel time from the starting point to the ending point. In the subsequent steps, operators of selection, crossover and mutation are iterated in order to improve the current population. The algorithm terminates after a fixed number of generations (denoted by N_g), or earlier if it converges. The current population is checked every 100 iterations and the algorithm is stopped if no improvements have been generated. The GA result is the route in the final population with the highest profit value. The steps of the GA are described in detail in the following subsections.

3.1 Initialization

The first step in using the GA to solve the OPTW is to code a solution (a route) into a chromosome. In our approach a route is coded as a sequence of locations. The length of the chromosome is not fixed because the number of locations on the route is not set. The parameter P_{size} denotes the size of the initial population. First the chromosome is initialized by the s and e vertices. Then the following values are assigned sequentially to the initialized vertices: $arrival_i$ - arrival time at vertex i, $wait_i$ - waiting time, if the arrival at a vertex i is before opening time, $start_i$ and end_i - starting and ending service time at vertex i. Moreover, the maximum time the service of a visit i can be delayed without making other visits infeasible is calculated for each location in the route as follows [16]:

$$MaxShift_i = Min(C_i - start_i - T_i, wait_{i+1} + MaxShift_{i+1}) \tag{1}$$

Let l be the predecessor of vertex e in the route. In the subsequent steps a set of vertices is prepared. Each vertex v from this set is adjacent to vertex l and vertex e and will satisfy the following conditions after insertion: (a) $start_v$ and end_v are within the range $[O_v, C_v]$; (b) the locations after v could be visited in the route; and (c) the current travel length does not exceed the given t_{max} (including consumption time to insert the vertex v between l and e). A random vertex v is chosen from this set. The values $arrival_v$, $wait_v$, $start_v$ and end_v are calculated and the vertex v is inserted. After the insertion, the values $arrival_e$, $wait_e$, $start_e$ and end_e are updated. Moreover, for each vertex in the tour (from vertex e to s) the $MaxShift$ value is updated as well. The tour generation is continued for as long as locations that have not been included are present and t_{max} is not exceeded.

3.2 Selection

We use tournament grouping selection, which yields better adapted individuals than standard tournament selection. We developed this method in a previous solution to the Orienteering Problem [7]. In this method a set of P_{size} individuals is divided into k groups and the tournaments are carried out sequentially in each of the groups. t_{size} random individuals are removed from the group, the chromosome with the highest value for the fitness function $TotalProfit^3/TravelTime$ is copied to the next population, and the t_{size} previously chosen individuals are returned to the old group (the power 3 in the fitness function was determined experimentally and the other values of the power gave worst results). After repetition of P_{size}/k selection from the group currently analysed, P_{size}/k individuals are chosen for a new population. Finally, when this step has been repeated in each of the remaining groups, a new population is created, containing P_{size} routes.

3.3 Crossover

In the crossover operator, first two random individuals are selected for the crossover stage. Then we determine all genes which could be replaced without exceeding the time window conditions and the t_{max} limit. We choose a set of genes with similar time windows and start and end of service. If there are no similar genes, crossover is terminated (no changes are applied). Otherwise, a random pair is selected from all similar pairs of genes. This pair is a point of crossover. Two new individuals are created by exchanging chromosome fragments (from the crossing point to the end of the chromosome) from both parents. Next, for each vertex i from the crossing point to vertex e, the values for $arrival_i$, $wait_i$, $start_i$ and end_i are updated and the new $MaxShift$ values are calculated for the locations from vertex e to s.

3.4 Mutation

In this phase of the GA a random route is selected from the P_{size} individuals. Two types of mutation are possible a gene insertion or gene removal (the probability of each is 0.5). The mutation process is repeated on the selected route N_m times, where N_m is the parameter of the GA. During the *insertion mutation*, all possibilities for inclusion of each new gene (not present in the chromosome) are considered. We check whether the shift resulting from the new insertion exceeds the constraints associated with the previously calculated $wait$ and $MaxShift$ values of the gene located directly after the newly inserted one. The location u with the highest value of $(p_u)^2/TravelTimeIncrease(u)$ is selected for insertion. $TravelTimeIncrease(u)$ is defined as the increased travel time after u is included. This value also takes into account the waiting and visiting time of vertex u. The selected gene u is inserted into the route and the values of $arrival_u$, $wait_u$, $start_u$ and end_u are calculated. For each location after u the arrival time, waiting time, and start and end of service are updated. Starting from the ending point, the $MaxShift$ value is updated for each location in the tour.

In the *deletion mutation* we remove a randomly selected gene (excluding the first and last genes) in order to shorten the travel length. After the gene is removed, all locations after the removed gene are shifted towards the beginning of the route. Furthermore, the locations before and after the removed gene should be updated as in the case of the insertion mutation.

4 Computational Experiment

The GA was implemented in C++ and run on an Intel Core i7, 1.73 GHz CPU (turbo boost to 2.93 GHz). The computational experiments were carried out on the well-known Solomon and Cordeau instances. It should be mentioned that Solomon's data set [13] and Cordeau's instances (pr01-pr10) [2] were adapted for the OPTW by Righini et al. [12]. These Solomon instances have 100 vertices: cluster class (c100), random class (r100) and random-clustered category (rc100). Moreover, Montemanni et al. [10] adapted 27 additional Solomon instances (c200, r200, rc200) and 10 instances based on Cordeau (pr11-pr20). The $c \setminus r \setminus rc200$ and the $c \setminus r \setminus rc100$ benchmarks have the same coordinates of vertices, profits and visiting times, but the $c \setminus r \setminus rc200$ instances have approximately three times higher values of t_{max} and proportionally larger time windows than the $c \setminus r \setminus rc100$ instances. The Cordeau instances vary between 48 and 288 vertices and t_{max} is equal to 1000. Several tests were carried out to establish the algorithm parameters and determine convergence and sensitivity. The parameter values which represent the best trade-off between the quality of the solutions and the computational time are as follows: P_{size} =150, t_{size}=3, k=15. The maximum number of iterations is 500, but every 100 generations the current population is checked and the GA is stopped if no improvements have been found. Moreover, during the testing we consider different parameters for the mutation repetition on the selected route: N_m= 3, 5, 10 and 15.

Tables 1 - 3 present detailed results of the GA performance with different numbers of mutations (the best score and the computational time in seconds). The GA was run 15 times; the best score and the total time of the 15 runs are given in the tables. The tables also show the percentage gap between the best solution values (BS) and the GA, and for comparison, the gap between BS and ILS as well. An empty cell denotes a gap equal to 0. The first eleven cells of the last rows of Tables 1 - 3 give the sums of the profits and times. The remaining cells show the average gap between BS and other results for the OPTW. The BS and the ILS results are taken from [4] and [16], respectively. Tests for which the GA improves the best known values are given in bold.

As shown in Table 1, for instances c100, r100 and rc100 (t_{max} is equal to 1,236, 230 and 240, respectively, and tight time windows are assigned to the locations) the GA provides a better solution on average than ILS even where only three mutations are used. For these instances, the average gap between BS and GA is only 0.37%. In contrast, the average gap between BS and ILS is 1.82%. Notably (see Table 2), for the instances c200, r200 and rc200 (t_{max} is equal to 3,390, 1,000 and 960, respectively, and wide time windows are assigned to the locations) our method gives similar score results to the ILS after application of ten or fifteen mutations. However, the execution time grows significantly for larger numbers of mutations. Table 3 presents the results obtained on Cordeau's data sets. In the pr11-20 instances wider time windows were given than in pr01-10. The GA outperforms the ILS in score results after five mutations, and the average gap between BS and GA decreases as the number of mutations increases. For the pr11 instance the GA improves the best known value.

Table 1. Results for Solomon's test problems (n=100)

			GA(number of mutations)								%gap with BS and				
			(3)		(5)		(10)		(15)						
name	ILS	BS	score	time	score	time	score	time	score	time	GA(3)	GA(5)	GA(10)	GA(15)	ILS
c101	320	320	320	3.72	320	5.18	320	8.22	320	10.61					
c102	360	360	360	4.65	360	6.47	360	9.08	360	11.42					
c103	390	400	390	5.03	390	6.5	400	11.47	400	14.99	2.5	2.5			2.5
c104	400	420	400	5.92	420	7.74	420	12.14	410	15.15	4.8			2.4	4.8
c105	340	340	340	4.42	340	5.48	340	8.18	340	9.5					
c106	340	340	340	3.59	340	4.37	340	7.19	340	9.92					
c107	360	370	360	4.3	370	5.63	370	8.63	370	10.51	2.7				2.7
c108	370	370	370	4.36	370	6.64	370	9.88	370	11.22					
c109	380	380	380	6.64	380	6.69	380	10.29	380	12.91					
r101	182	198	198	2.7	198	3.55	198	5.92	198	7.92					8.1
r102	286	286	286	4.36	286	6.14	286	9.67	286	12.57					
r103	286	293	293	5.02	293	5.72	293	9.91	293	12.69					2.4
r104	297	303	298	5.02	303	6.18	303	9.63	303	14.12	1.7				2.0
r105	247	247	247	3.19	247	4.61	247	7.72	247	11.31					
r106	293	293	293	4.23	293	6.35	293	8.9	293	10.99					
r107	288	299	297	4.4	299	6.82	299	12.96	297	13.7	0.7			0.7	3.7
r108	297	308	301	4.87	308	6.32	308	12.19	308	15.57	2.3				3.6
r109	276	277	276	3.35	274	5.32	276	10.14	276	13.05	0.04	0.01	0.04	0.04	0.04
r110	281	284	281	4.04	281	5.52	281	9.86	281	13.79	1.1	1.1	1.1	1.1	1.1
r111	295	297	295	4.78	295	5.94	297	11.68	295	16.1	0.7	0.7		0.7	0.7
r112	295	298	295	4.96	298	7.12	295	11.44	295	15.56	1.0	1.0	1.0	1.0	1.0
rc101	219	219	219	2.68	219	3.94	219	7.21	219	10.02					
rc102	259	266	266	3.25	266	5.54	266	9.1	266	11.06					2.6
rc103	265	266	262	3.33	266	5	266	9.03	266	11.59	1.5				0.4
rc104	297	301	301	3.91	297	5.9	301	9.91	301	15		1.3			1.3
rc105	221	244	241	3.02	244	4.43	244	9.42	244	11.48	1.2				9.4
rc106	239	252	250	3.61	250	4.8	250	8.17	250	10.85	0.8	0.8	0.8	0.8	5.2
rc107	274	277	276	4.08	274	6.06	277	9.74	277	12.47	0.4	1.1			1.1
rc108	288	298	298	4.49	298	5.87	298	9.24	298	11.35					3.4
	8645	8806	8773	121.92	8779	165.83	8797	277.16	8783	358.1	0.37	0.3	0.1	0.26	1.82

Table 2. Results for Solomon's test problems, cont. (n=100)

			GA(number of mutations)								%gap with BS and				
			(3)		(5)		(10)		(15)						
name	ILS	BS	score	time	score	time	score	time	score	time	GA(3)	GA(5)	GA(10)	GA(15)	ILS
c201	840	870	850	7.83	850	16.88	860	26.76	860	37.87	2.3	2.3	1.1	1.1	3.4
c202	910	930	870	14.28	910	23.25	920	34.81	920	45.08	6.5	2.2	1.1	1.1	2.2
c203	940	960	900	16.76	930	25.69	940	46.76	940	57.63	6.3	3.1	2.1	2.1	2.1
c204	950	980	920	21.16	940	28.9	960	53.75	970	66.95	6.1	4.1	2.0	1.0	3.1
c205	900	910	880	9.71	900	23.61	900	30.71	900	41.82	3.3	1.1	1.1	1.1	1.1
c206	910	930	900	16.8	910	20.71	910	36.44	920	42.79	3.2	2.2	2.2	1.1	2.2
c207	910	930	900	16.83	910	19.2	920	38.08	920	45.83	3.2	2.2	1.1	1.1	2.2
c208	930	950	930	15.98	920	23.49	940	39.24	940	42.52	2.1	3.2	1.1	1.1	2.1
r201	788	797	778	11.64	760	23.29	775	45.88	780	65.15	2.4	4.6	2.8	2.1	1.1
r202	880	929	9863	18.63	897	33.89	878	55.19	886	67.45	7.1	3.4	5.5	4.6	5.3
r203	980	1021	895	21.68	918	33.14	958	73.3	989	94.64	12.3	10.1	6.2	3.1	4.0
r204	1073	1086	985	26.15	1011	43.68	1034	79.99	1047	106.18	9.3	6.9	4.8	3.6	1.2
r205	931	953	846	15.02	875	26.07	933	57.34	920	73.6	11.2	8.2	2.1	3.5	2.3
r206	996	1029	922	22.75	935	30.72	1002	68.57	994	86.58	10.4	9.1	2.6	3.4	3.2
r207	1038	1072	967	30.13	992	42.26	991	70.04	1035	104.84	9.8	7.5	7.6	3.5	3.2
r208	1069	1112	995	27.1	1045	45.84	1059	91.87	1065	109.03	10.5	6.0	4.8	4.2	3.9
r209	926	950	875	23.66	913	35.46	914	63.12	916	83.1	7.9	3.9	3.8	3.6	2.5
r210	958	987	874	20.68	925	40.2	945	56.7	942	92.74	11.4	6.3	4.3	4.6	2.9
r211	1023	1046	953	27.28	971	40.95	1006	69.08	1013	106.34	8.9	7.2	3.8	3.2	2.2
rc201	780	795	739	9.24	757	18.08	777	31.17	788	49.37	7.0	4.8	2.3	0.09	1.9
rc202	882	936	855	14.56	898	29.47	881	43.64	908	65.93	8.7	4.1	5.9	3.0	5.8
rc203	960	1003	921	19.86	851	26.83	942	55.22	958	65.03	8.2	5.2	6.1	4.5	4.3
rc204	1117	1136	1032	23.84	1026	26.92	1088	52.45	1074	77.06	9.2	9.7	4.2	5.5	1.7
rc205	840	859	803	9.29	808	18.94	842	41.9	832	63.26	6.5	5.9	2.0	3.1	2.2
rc206	860	895	840	12.05	850	21.76	856	37.34	860	53.9	6.1	5.0	4.4	3.9	3.9
rc207	926	983	878	16.89	911	29.39	932	48.96	965	66,92	10.7	7.3	5.2	1.8	5.8
rc208	1037	1053	980	19.43	1013	31.39	1033	59.17	1014	65.43	6.9	3.8	1.9	3.7	1.5
	25354	26102	24151	489.23	24817	779.981	25196	1407.48	25356	1877.04	7.47	4.92	3.47	2.86	2.86

Table 3. Results for Cordeau's test problems (n from 48 to 288)

			GA(number of mutations)								%gap with BS and				
			(3)		(5)		(10)		(15)						
n name	ILS	BS	score	time	score	time	score	time	score	time	GA(3)	GA(5)	GA(10)	GA(15)	ILS
48 pr01	304	308	304	4.3	302	5.18	308	9.36	305	13.68	1.3	1.9		1.0	1.3
96 pr02	385	404	373	6.18	394	10.68	393	19.48	396	25.32	7.2	2.5	2.7	2.0	4.7
144 pr03	384	394	373	8.28	384	15.29	384	25.35	386	33.67	5.3	2.5	2.5	2.0	2.5
192 pr04	447	489	438	11.99	447	21.81	470	44.29	466	56.57	10.4	8.6	3.9	4.7	8.6
240 pr05	576	595	505	15.51	560	32.88	568	61.87	581	79.01	15.1	5.9	4.5	2.4	3.2
288 pr06	538	590	533	18.79	556	30.99	560	64.88	555	90.85	9.7	5.8	5.1	5.9	8.8
72 pr07	291	298	291	3.52	298	6.4	293	10.67	291	15.76	2.3		1.7	2.3	2.3
144 pr08	463	463	419	7.99	437	16.32	448	28.53	444	37.89	9.5	5.6	3.2	4.1	
216 pr09	461	493	444	12.79	470	23.7	457	41.26	477	70.11	9.9	4.7	7.3	3.2	6.5
288 pr10	539	594	516	19.59	550	38.06	537	56.82	566	108.08	13.1	7.4	9.6	4.7	9.3
48 **pr11**	**330**	**330**	**330**	3.74	**331**	6.39	**346**	11.65	**342**	19.8		-0.03	-4.8	-3.9	
96 pr12	431	442	417	10.36	420	11.13	436	24.56	432	36.88	5.7	5.0	1.4	2.3	2.4
144 pr13	450	461	437	10.93	443	14.53	454	35.97	450	53.82	5.2	3.9	1.5	2.4	2.4
192 pr14	482	567	484	15.46	508	26.08	505	47.96	512	72.01	14.6	10.4	10.9	9.7	15.0
240 pr15	638	685	625	20.02	630	37.74	648	75.381	664	123.36	8.8	8.0	5.4	3.1	6.9
288 pr16	559	674	559	20.12	595	42.35	598	77.94	611	121.46	17.1	11.7	11.3	9.3	17.1
72 pr17	346	359	343	5.72	344	8.74	356	16.69	356	19.5	4.5	4.2	0.8	0.8	3.60
144 pr18	479	535	473	12.28	507	19.5	498	33.24	521	51.46	11.6	5.2	6.9	2.6	10.5
216 pr19	499	562	475	16.01	498	28.12	530	61.97	533	79.29	15.5	11.4	5.7	5.2	11.2
288 pr20	570	667	561	19.75	580	37.62	619	89.77	615	147.58	15.9	13.0	7.2	7.8	14.5
	9172	9910	8902	243.33	9254	433.51	9408	837.64	9503	1255.48	10.17	6.61	5.06	4.01	7.44

5 Conclusions and Further Work

The main contribution of this paper is a new genetic algorithm which additionally exploits elements of the local search method to solve the orienteering problem with time windows. The computational experiments show that repetition of the mutation step for the same route during one GA iteration can improve the solution, outperforming the ILS score result. While ILS is very fast [16], the GA execution could easily be divided up and executed on multiple processors. In the experiment described in this article, the best score and execution time from 15 runs of the GA were taken into account. Each of the 15 runs could be executed on a separate processor, decreasing the time significantly. Furthermore, the grouping selection could be executed in parallel in each group.

In our further research it is our intention to conduct experiments on large networks of locations with realistic time windows. We will use a base of our city's POI with realistic opening and closing time windows, as well as ratings by internet users. A typical trip takes a few hours and most of the POI may be open all day, so time windows are very wide and could overlap. Adding some improvements to the algorithm and testing this particular case will be a crucial issue.

Acknowledgements. The authors gratefully acknowledge support from the Polish Ministry of Science and Higher Education at the Bialystok University of Technology (grant S/WI/1/2011 and W/WI/2/2013).

References

1. Archetti, C., Hertz, A., Speranza, M.G.: Metaheuristics for the team orienteering problem. Journal of Heuristics 13, 49–76 (2007)
2. Cordeau, J.F., Gendreau, M., Laporte, G.: A tabu search heuristic for periodic and multi-depot vehicle routing problems. Networks 30(2), 105–119 (1997)
3. Garcia, A., Vansteenwegen, P., Arbelaitz, O., Souffriau, W., Linaz, M.: Integrating Public Transportation in Personalised Electronic Tourist Guides. Computers & Operations Research 40, 758–774 (2013)
4. Labadi, N., Mansini, R., Melechovsky, J., Calvo, R.W.: The Team Orienteering Problem with Time Windows: An LP-based Granular Variable Neighborhood Search. European Journal of Operational Research 220(1), 15–27 (2012)
5. Kantor, M., Rosenwein, M.: The Orienteering Problem with Time Windows. Journal of the Operational Research Society 43, 629–635 (1992)
6. Karbowska-Chilinska, J., Koszelew, J., Ostrowski, K., Zabielski, P.: Genetic algorithm solving orienteering problem in large networks. Frontiers in Artificial Intelligence and Applications 243, 28–38 (2012)
7. Karbowska-Chilinska, J., Koszelew, J., Ostrowski, K., Zabielski, P.: A Genetic Algorithm with Grouping Selection and Searching Operators for the Orienteering Problem (under review)
8. Karbowska-Chilińska, J., Zabielski, P.: A Genetic Algorithm vs. Local Search Methods for Solving the Orienteering Problem in Large Networks. In: Graña, M., Toro, C., Howlett, R.J., Jain, L.C. (eds.) KES 2012. LNCS, vol. 7828, pp. 11–20. Springer, Heidelberg (2013)

9. Ostrowski, K., Koszelew, J.: The comparison of genetic algorithm which solve Orienteering Problem using complete an incomplete graph. Zeszyty Naukowe, Politechnika Bialostocka, Informatyka 8, 61–77 (2011)
10. Montemanni, R., Gambardella, L.M.: Ant colony system for team orienteering problems with time windows. Foundations of Computing and Decision Sciences 34 (2009)
11. Righini, G., Salani, M.: New dynamic programming algorithms for the resource constrained elementary shortest path. Networks 51(3), 155–170 (2008)
12. Righini, G., Salani, M.: Dynamic programming for the orienteering problem with time windows. Technical Report 91, Dipartimento di Tecnologie dell'Informazione, Universita degli Studi Milano, Crema, Italy (2006)
13. Solomon, M.: Algorithms for the Vehicle Routing and Scheduling Problems with Time Window Constraints. Operations Research 35(2), 254–265 (1987)
14. Tsiligirides, T.: Heuristic methods applied to orienteering. Journal of the Operational Research Society 35(9), 797–809 (1984)
15. Vansteenwegen, P., Souffriau, W., Vanden Berghe, G., Van Oudheusden, D.: The City Trip Planner: An expert system for tourists. Expert Systems with Applications 38(6), 6540–6546 (2011)
16. Vansteenwegen, P., Souffriau, W., Vanden Berghe, G., Van Oudheusden, D.: Iterated local search for the team orienteering problem with time windows. Computers O.R. 36, 3281–3290 (2009)

The Study on the Collusion Mechanism of Auction and It's Efficiency

Huirong Jing

School of Management, Yunnan University of Nationalities, Chengong,
Kunming, Yunnan, P.R. China
1307608193@qq.com, gouhuirong@yahoo.com.cn

Abstract. This paper concerns the cartel with payment of monetary compensation, studies the cartel members' bidding strategies in the pre-auction under the first-price and second-price collusion mechanism. It is found that the cartel members report price below its valuation in the first-price collusion mechanism, the cartel members' offer price is higher than its valuation in the second price collusion mechanism. Furthermore, study the first-price and second-price collusion mechanism's efficiency when the numbers of cartel members is less than and equal to the total numbers of bidders.

Keywords: auction, cartel, collusion, efficiency.

1 Introduction

Auction collusion between bidders is widespread, it is not only in sealed bid auction, for example: [1] found that there existed collusion in auction through detecting U.S. Highway construction bid contracts, [16] found that school milk distribution businesses colluded to manipulate the market in Fuluolida State and the Texas State, [8,11] studied the collusion in auction of the Dutch construction market construction companies from different angles. But also auction collusion is in the English auction, such as [4] studied collusion among bidders in the auction of the right to deforestation in U.S.

The influence of the auction mechanism on the efficiency of collusion has been documented in a number of empirical works. [1] studied the impact of internal decision-making structures on the collusion stability. To this end, they established a three-firm spatial competition model where two firms belong to the same holding company. The holding company can decide to set prices itself or to delegate this decision to its local units. It is shown that when transportation costs are high, collusion is more stable under delegation. Furthermore, collusion with maximum prices is more profitable if price setting is delegated to the local units. [6] shown that a tax rate which depended on the pollution stock, can induce stable cartelization in an oligopolistic polluting industry.

The degree of auction collusion can be directly related to the payment of monetary compensation and redistribute the object, the study of the efficiency of this auction

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_60, © Springer International Publishing Switzerland 2014

collusion mechanism also attracted the attentions of scholars. [14] shown that the extent of collusion was tied to the availability of transfers. Monetary transfers allow cartels to extract full surplus under the implied condition of bidders could report its true valuation to pre-auction mechanism and the condition of auctioneer announced the reserve price.

In [9] authors studied collusion in an IPV auction with binary type spaces. Collusion is organized by a third party that can manipulate participation decisions. It shown that collusion in the optimal auction is efficient when the third party can implement monetary transfers as well as when it can implement monetary transfers and reallocations of the object. The threat of non-participation in the auction by a subset of bidders is crucial in constraining the seller's profit.

In [2] they studied collusion in repeated auctions when bidders communicate prior to each stage auction. For independent and correlated private signals and general interdependent values, they identified conditions under which an equilibrium collusion scheme is fully efficient in the sense that the bidders' payoff is close to what they got when the object was allocated to the highest valuation bidder at the reserve price in every period.

In [3] they presented a simple dynamic bid rotation scheme which coordinates bids based on communication history and enables inter temporal transfer of bidders' payoffs. It derived a sufficient condition for such a dynamic scheme to be an equilibrium and characterizes the equilibrium payoffs in a general environment with affiliated signals and private or interdependent values. With IPV, it was shown that the dynamic scheme yields a strictly higher payoff to the bidders than any static collusion scheme which coordinates bids based only on the current reported signals.

Collusion among bidders damages the benefit of the auctioneer, to the perspective of revealing collusion feature,[13] compared the collusive properties of two standard auctions, the English auction and the first-price sealed bid auction, and a lesser known format, the Amsterdam (second-price) auction. It was study two settings: in one, all bidders can collude, and in another, only a subset was eligible. The experiments shown that the Amsterdam auction triggers less collusion than the standard auctions.

Based on the point-in-time after the bid-rigging success of cartel collusion, [10] studied the losses caused by the collusion to the auctioneer from the bidders' bidding strategies in monetary compensation cartel knockout auction. They found the bidders' bid was not independent of the sharing rule of knockout auction, and the win price of cartel and the total payment of winning bidder was not the value of unbiased estimation in the non-collusive auction, and estimated its deviation range from non-collusion.

Authors of [17] implemented experimental methods to study that the effect of the rules of improve in ascending price auction and the asymmetric bidder's information about estimation value on the bidders collusion, he found it can not always destroy the collusion of the bidders with strong speculative desire when the strict tender rules improve and the bidder's private valuation information change.

Under the condition of bidders' valuation ware discrete allocated, [7] studied the collusive equilibrium problems between two bidders with unilateral monetary compensation payment when the non-collusion number was a random number in single second-price auction from the perspective of collusive members may deceive each other. They found that the unilateral monetary compensation payment could successfully guide the conspiracy, and the mastermind advocator's income will increase.

Authors of [12] studied the bidders' collusion of the second-price and English auction, designe the second-price collusion mechanism of which the cartel with the payment of monetary compensation divided the collusive gains which is the difference between the second highest price in its pre-auction and obtaining the object costs in public auction equally to the cartel members. They pointed out that under this collusive mechanism, each cartel member will report its valuation truely, bidders participating in the collusion is a dominant strategy, equilibrium involving in the collusion number is equal to the number of the bidders, and pointed out that the collusive bidders held a pre-auction before the public auction can complete allocation of benefits rational.

Authors of [5] analyzed bidder collusion from the two aspects whether the cartel could control the bids of their members. They found that cartels that cannot control the bids of their members can eliminate all ring competition at second-price auctions, but not at first-price auctions. At first-price auctions, when the cartel cannot control members' bids, cartel behavior involves multiple cartel bids. Cartels that can control bids of their members can suppress all ring competition at both second-price and first-price auctions; however, shill bidding reduces the profitability of collusion at first-price auctions.

Based on the point-in-time before the cartel collusion, this paper borrows [10]'s research methods about the bidders bid strategies in cartel knockout auction. Through establishing the theory model of the first-price and second-price collusion mechanism of cartel's pre-auction and detecting [14] proposed the implicit assumption and the assumption proposed by [12] which the second-price collusion mechanism could guide the members of the cartel real quote, it studys the cartel members' bidding strategy in the first-price collusive mechanism and the second-price collusive mechanism respectively proposed by [12,14]. Moreover, when the number of cartel members is less than the total number of bidders and equal to the total number of bidders, we further analysis bidding strategy of catel representatives participating in the first-price and second-price public auction rather than [14]'s strategic distribution from the research perspective different from literature [5] and different assumptions of [7], modify the [18]about the bidders' valuation distribution.

Under the condition of the bidders' value is continuous and is uniform distribution, we study the collusion problem in the one-shot first-price and second-price auction, analysis that the cartel held the first-price and second-price collusive mechanism pre-auction internally before the first-price and second-price public auction to determine the representative to participate the public auction and the way to distribution the benefits of collusion. Furthermore, through investigating the cartel members' bidding strategies in the two collusion mechanism, and the bidding strategies in participating in the first-price and second-price sealed bid public auction when the number of collusive members is less than and equal to total number of bidders, we study whether the auction collusive benefit could be achieved, thus to reveal the auction collusive efficiency issues, and obtain study results different from the [13].

2 Basic Assumptions

This Paper assumes that the cartel could held a pre-auction before public auction, and there is a pre-auction center, each cartel member first of all reports his bid to the pre-auction center, and the centre decides the representative to participate in the public auction according to its offer and the way of distributing collusive benefits. Main study is the cartel's first-price and second-price collusive benefit allocation mechanism.

Cartel's first-price collusive benefit allocation mechanism is that the first-price sealed bid auction is held internally among cartel members before the public auction, and the highest bidder on behalf of the cartel participates the public auction, if the representative of cartel gets the object in the public auction and the cost of the object in the public auction is higher than the first highest price in the pre-auction , the representative of the cartel takes out the collusive benefit which is the difference between the highest price (collusive Price) in the pre-auction and the cost of object in public auction, and divides the collusive benefit equally to the members of the cartel.

Cartel's second-price collusive benefit allocation mechanism is that the second-price sealed bid auction is held internally among cartel members before the public auction, and the highest bidder on behalf of the cartel participates the public auction, if the representative of cartel gets the object in the public auction and the cost of the object in the public auction is higher than the second highest price in the pre-auction, the representative of the cartel takes out the collusive benefit which is the difference between the second highest price (collusive Price) in the pre-auction and the cost of object in public auction, and divides the collusive benefit equally to the members of the cartel.

Bidders and auctioneer are all risk-neutral; The total number of bidder is n, The number of collusive bidder is $k, 2 \le k \le n$; bidders' valuation of object is independent and identically distributed, and obeys uniform distribution of $[0,1]$, the bidders are symmetric homogeneous and the valuation $v_i \in [0,1]$, $i = 1, 2...n$; The same bidder has the same valuation about the object either in public auction or in pre-auction of the cartel, The bidder offers $b(v_i)$ $i = 1, 2...n$ as his valuation is v_i , and the bid is strictly increasing continuous function, that the valuation high bidders offer higher bid; And assume that the bidders do not report their real valuations in the pre-auction.

Further assume that the costs of the cartel representative to obtain object in the public auction is c_a ; The collusive price is p_c ,when cartel uses the first price collusive benefits allocation mechanism, p_c means that the first highest price in the pre-auction; when cartel use the second price collusive benefits allocation mechanism, p_c means that the second highest price in the pre-auction.

3 The Cartel's Collusive Pre-auction and It's Members' Bidding Strategies

The Symmetric homogeneity of bidders determines that the competitive bidding will reduce the bidder's benefits, collusion among bidders could increase bidder's profits, and thus bidding collusion has potential power. Firstly, we consider the cartel members' bidding strategies in the pre-auction when the cartel uses the first-price collusive benefit allocation mechanism. As the highest bidder in the pre-auction will be the representative of the cartel, if the collusion is successful, namely, the highest price in the pre-auction is higher than the cost of the representative to get the object in the public auction, the representative of the cartel takes out the collusive benefit which is the difference between the highest price (collusive Price) in the pre-auction and the cost of object in public auction, and divides the collusive benefit equally to the members of the cartel.

Under the first-price collusive benefits allocation mechanism, the cartel member reports t_i as his value is v_i and offers $b(t_i)$, $i = 1, 2...k$, his income is:

$$E(t_i, v_i) = [v_i - b(t_i) + \frac{1}{k}(b(t_i) - c_a)]t_i^{k-1} + \int_{t_i}^{\bar{v}} \frac{1}{k}[b(z) - c_a][k-1] * 1 * z^{k-2} dz \tag{1}$$

The first item is the cartel members' expected income which his bid is the highest when he report t_i as his valuation is v_i , the second item is his expected income when his reporting price living in the second high-priced and below.

Let $E(t_i, v_i)$ derivate to t_i .and let it is 0, by the envelope theorem, we can know:

$$\left. \frac{\partial E(t_i, v_i)}{\partial t_i} \right|_{t_i = v_i} = 0 \tag{2}$$

That is:

$$(k-1)v_i^{k-2}\left[v_i - \frac{k-1}{k}b(v_i) - \frac{1}{k}c_a\right] - \frac{k-1}{k}v_i^{k-1}b'(v_i) - \frac{k-1}{k}[b(v_i) - c_a]v_i^{k-2} = 0 \tag{3}$$

Simplification and finishing , the times on the both sides of the equation with v_i ,then there is:

$$\frac{d\left[v_i^k b(v_i)\right]}{dv_i} = kv_i^k \tag{4}$$

When $v_i = 0$, $\left[v_i^k \times b(v_i)\right] = 0$, Solving the differential equation to meet this initial value, there is:

$$b(v_i) = \frac{k}{k+1}v_i < v_i, i = 1, 2...k \text{。} \tag{5}$$

Under the first-price collusive benefit allocation mechanism, the cartel members' offer is independent of the costs to get the object in public auction and less than his valuation. Similarly, the first-price collusive benefits allocation mechanism does not guide the cartel members to truly report their valuations. It is because that the personal interests' objective of the members of the cartel is not completely consistent with target of the cartel. When the costs of getting the object in public auction are certain, lower pricing could be reduced revenue to give the other cartel members, and increase revenue as a representative of the cartel.Then from the equation (5), we can get the following proposition.

Proposition 1: *In the first-price collusive benefit allocation mechanism pre-auction, the cartel member's reporting price is less than his valuation.*

Secondly, we consider the cartel members' bidding strategies in the pre-auction when the cartel uses the second-price collusive benefit allocation mechanism. The highest bidder in the second-price sealed pre-auction will be the representative of the cartel to participate the public auction, if the collusion is successful, the representative of the cartel takes out the collusive benefit which is the difference between the second highest price (collusive Price) in the pre-auction and the cost of object in public auction, and divides the collusive benefit equally to the members of the cartel. Under the collusive mechanism, the cartel member reports t_i as his value is v_i and offers $b(t_i)$, $i = 1, 2...k$, his expected income is:

$$E_2(t_i, v_i) = \int_{\underline{v}}^{t_i} [v - b(z) + \frac{1}{k}(b(z) - c_a)](k-1) * 1 * z^{k-2} dz + (k-1) t_i^{k-2} \left[1 - t_i\right] \frac{1}{k}(b(t_i) - c_a)]$$
$$+ \int_{t_i}^{\bar{v}} \frac{1}{k}[b(z) - c_a](k-1)(k-2) * 1 * z^{k-3} \left[1 - z\right] dz \tag{6}$$

The first item and the second item are respectively the cartel members' expected income being the first and second high-priced when he report t_i as his valuation is v_i , the third item is his expected income when his reporting price living in third high-priced and below.

Similar to the previous method, Let $E_2(t_i, v_i)$ derivate to t_i , and let it is 0,By the envelope theorem such as equation (2), Simplification finishing, Then:

$$\frac{d[(1 - v_i)^k b(v_i)]}{dv_i} = v_i \frac{d(1 - v_i)^k}{dv_i} \tag{7}$$

When $v_i = 1$, have $(1 - v_i)^k b(v_i) = 0$, solving the differential equation to meet this initial value, there is:

$$b(v_i) = v_i + \frac{1}{k+1}(1 - v_i), i = 1.2..k \tag{8}$$

And must include:

$$b(v_i) = v_i + \frac{1}{k+1}(1 - v_i) > v_i, i = 1.2..k \tag{9}$$

Otherwise,

$$v_i = 1, b(v_i) = 1, i = 1.2..k \tag{10}$$

It is a contradiction with previous assumptions which the bidder's bid is strictly monotonous increasing continuous function.

Namely, under the second-price collusive benefit allocation mechanism, every cartel member does not report price truly, his reporting price is higher than his valuation. And his reporting price is also independent of the costs to get the object in public auction. It is the same due to the personal interest's objectives of the cartel members are not completely consistent with target of the cartel. High offer may increase the probability to become the representative of the cartel, thereby increasing his income. Then we can get the following proposition.

Proposition 2: cartel members' reporting price depends on the collusive mechanism, in the second-price collusive benefit allocation mechanism pre-auction, the cartel members' reporting price is higher than his valuation.

The cartel held the pre-auction internally, the cartel members' reporting price is different in different allocation mechanisms. It shows that: the cartel members' actions are speculative, and their objectives are to maximize their incomes, and thus, it is difficult to demand the members of the cartel acting complied with the supreme principle of the cartel's whole benefits.

4 Cartel Members' Bidding Strategy in the Public Auction and Collusive Efficiency

Because the bidders' bidding strategy is equivalent in Dutch auction and the first-price sealed auction, and the bidders' bidding strategy in the English auction is equivalent to the second-price sealed auction. The public auction is only studied in the first-price and second-price sealed bid auctions in this paper.

When k = n, that is to say, the number of cartel members is equal to the total number of bidders. If Public auction is the first-price sealed bid auction and the reserve price is announced, the representative of the cartel directly bids the reserve price, the cost of obtaining the object is the lowest, it is optimal strategy for the representative per se and the cartel, at this time, the cartel representative's bid is independent of collusive sharing mechanism.

The speculativeness of the cartel members' actions makes the other cartel members have the motivation to quote price slightly higher than the reserve price, and get more profits. If the auctioneer hides the reserve price, the cartel needs firstly to calculate the reserve price, followed it needs to organize its members to bid collaboratively, namely, the cartel representative reports the specified high-price, the other members of cartel report corresponding low-price collaboratively. The cartel representative must strictly stick to the principle of collusive bidding successfully and getting the object at the lowest cost to gain his optimal benefits. The other members of the cartel have more detail information about the collusive bidding quote price, as long as the income of destruction of collusion is higher than the profit of collusion , the other cartel members have an incentive to quote slightly higher than the cartel representative to obtain the object, and get more benefits by destruction of collusion. Therefore, we can see that the first-price public auction is not conducive to the stability of the cartel collusion, hidden reserve price adds the difficulty of the cartel collusion.

Whereas the public auction is second-price sealed bid auction, for whether the reserve price is announced, taking into account the speculative actions of the cartel members, namely, the possibility of constructing collusion, the cartel representative reports any high-price, even the sky-high price, the cost of obtaining the object is independent of his offer, and it is optimal strategy for the representative per se and the cartels.

Now the cartel representative's bid is independent of the collusive benefit-sharing mechanisms, the other members of the cartel are more willing to maintain the collusion at corresponding low price, otherwise, the higher offers of the other cartel members will raise the cartel's costs to obtain the object, reduce collusive income, and the cartel members' income will be reduced with the reduction of the collusive income; or he has a chance to get the "winner's curse" (win-object unprofitably). Thus, when k = n, the second-price public auction is more conducive to the success of the cartel collusion. This may also explain why in reality people prefer using first-price auction, such as in antiques, art auction, procurement auction, TV advertising rights auction, even electricity and other franchised auction, they are all employing first-price auction. Then we have the following proposition.

Proposition 3: *When k = n, that is, all bidders are involved in the collusion, the public auction is the second-price sealed bid auction, the collusion is more stable, and the first-price sealed bid auction and hiding the reserve price in the public auction will increase the difficulty of cartel collusion.*

When k < n, similarly, when the number of collusive bidders is less than the totel number of bidders and the auctioneer does not announce the reserve price, the other members of cartel have no incentive to break the collusion, at this time, considering the cartel representative participating in the public auction, and the number of effective bidders participating in the public auction is $n-k+1$.

In the first-price sealed-bid public auction, The cartel representative select bidding strategy $b(v_i)$ to maximize his expected return.

$$E(f)=\left[(v_i-p_c)+\frac{1}{k}(p_c-b(v_i))\right]v_i^{n-k} \tag{11}$$

Let $E(f)$ derivate to v_i, and let it is 0. Simplification finishing:

$$\frac{d[\frac{1}{k}b(v_i)\times v_i^{n-k}]}{dv_i}=v_i^{n-k}+(v_i-p_c+\frac{1}{k}p_c)[v_i^{n-k}]' \tag{12}$$

Meet: when $v_i=0$,

$$\frac{1}{k}b(v_i)\times v_i^{n-k} = 0 \tag{13}$$

Solving differential equations to meet this initial value, we can obtain:

$$b(v_i) = k(v_i - p_c) + p_c \tag{14}$$

If the cartel use the first-price collusion mechanism, From Proposition 1 it is known that: $v_i - p_c > 0$,then: $b(v_i) > p_c$

If the cartel uses the first-price collusion mechanism, by Proposition 1 we know that: $v_i - p_c > 0$,then: $b(v_i) > p_c$,

That is to say the bid of the cartel representative is greater than the first price in the pre-auction, the cost of cartel to get the object is too high in the public auction and it is higher than the collusive price, the cartel has no collusive benefits to divide to the cartel members, and the collusion failed. If the cartel uses the second-price collusive mechanism, when $v_i - p_c > 0$,that is to say the cartel representative's valuation is higher than the collusive price, there is $b(v_i) > p_c$,

namely, the cost of cartel to get the object in the public auction is higher than the collusive price, the cartel has no collusive benefits to divide to the cartel members. The collusion is non-efficiency.

In the second-price sealed bid public auction, the cartel representative report t_i as his valuation is v_i and bid $b(v_i)$ to maximize his expected return,

$$\begin{aligned} &E(C) = \left[(v_i - p_c) + \frac{1}{k}(p_c - b(y)) \right] \Pr ob\left\{ b(y_j) < b(v_i) \middle| j = 1...i-1, i+1...n-k+1 \right\}, \\ &y = \max\left\{ v_1 ... v_{i-1,}\, v_{i+1} ... v_{n-k+1} \right\} \end{aligned} \tag{15}$$

That is:

$$E(c) = \int_{\underline{v}}^{t} [(v_i - p_c) + \frac{1}{k}(p_c - b(y))](n-k)y^{n-k-1} * 1dy \tag{16}$$

By the envelope theorem such as equation (2) ,That is:

$$b(v_i) = k(v_i - p_c) + p_c \tag{17}$$

If the cartel uses the first-price collusive mechanism, similar to the previous case, the cartel representative's bid is higher than the first price in the pre-auction, the collusion fails. If the cartel uses the second-price collusive mechanism, when $v_i - p_c > 0$, that is to say the cartel representative's valuation of object is higher than the collusive price, $b(v_i) > p_c$, the cost of cartel to get the object in the public auction is higher than the collusive price, the cartel havs no collusive benefits to divide to the cartel members, the collusion fails. It can be concluded:

Proposition 4: k<n, the auctioneer uses any form of hidden reserve price in public auction, if the cartel uses the first-price collusive mechanism, the collusion is non-efficiency. If the cartel uses the second-price collusive mechanism, when the cartel

representative's valuation of object is higher than the collusive price, the collusion is non- efficiency.

5 Conclusion

When the bidders are symmetric homogeneous, bidders collusion could increase the income of the bidders with respect to competitive bidding, which determines the bidders collusion has potential power. However, the bidders collusion not only harms the interests of the auctioneer, but also damages social resource allocation efficiency. This paper studies cartel members' bidding strategy in the pre-auction and public auction under the two typical cartel collusion mechanism which has monetary compensation payment, to reveal the efficiency of the auction collusion.

This Paper studies cartel members' bidding strategy in the pre-auction and public auction under the two typical cartel collusion mechanism which has monetary compensation payment, to reveal the efficiency of the auction collusion. It is found that cartel members' report price depends on the collusive mechanism. In the first-price collusive benefit allocation mechanism, the cartel members' reporting price is less than their valuation. And in the second-price collusive benefit allocation mechanism, the cartel members' reporting price is higher than his valuation. Furthermore, we found that, when the number of cartel members is equal to the numbers of the total number of bidders, the second-price sealed bid public auction is more conducive to the stability of the cartel collusion than the first-price sealed bid auction, and in the first-price sealed bid auction, hidden reserve price will increase difficulty of collusion. When the number of cartel members is less than the total number of bidders, the auctioneer uses any form of auction with hidden reserve price in the public auction, if the cartel uses the first-price collusive mechanism, the collusion is non- efficiency. If the cartel uses the second-price collusive mechanism, when the cartel representative's valuation of object is higher than the collusive price, the collusion is non- efficiency.

In order to ensure the efficiency of the resources allocation and the auctioneer's benefits, the auctioneer uses the first-price sealed bid auction hidden reserve price, and expands the scope of the bidders from the technical form to shorten the time between the information published and the auction, such as using the characteristics of the Internet's offsite and fastness, increasing the scope of the bidders, trying to prevent the information exchanges between the bidders and blocking all bidders to participate in the collusion, it could improve the resource allocation efficiency of auction.

References

1. Rasch, A., Wambach, A.: Internal decision-making rules and collusion. Journal of Economic Behavior & Organization 72, 703–715 (2009)
2. Aoyagi, M.: Efficient collusion in repeated auctions with communication. Journal of Economic Theory 134, 61–92 (2007)

3. Aoyagi, M.: Bid rotation and collusion in repeated auctions. Journal of Economic Theory 112, 79–105 (2003)
4. Baldwin, L., Marshall, R., Richard, J.F.: Bidder collusion at forest service timber sales. Journal of Political Economy 105(4), 657–699 (1997)
5. Marshall, R.C., Marx, L.M.: Bidder collusion. Journal of Economic Theory 133, 374–402 (2007)
6. Benchekroun, H., Chaudhuri, A.R.: Environmental policy and collusion: The case of a dynamic polluting oligopoly. Journal of Economic Dynamic & Control 35, 479–490 (2011)
7. Chen, C.-L., Tauman, Y.: Collusion in one-shot second-price auctions. Economic Theory 28, 145–172 (2006)
8. Doree, A.G.: Collusion in the Dutch construction industry: an industrial organization perspective. Building Research & Information 32(2), 146–156 (2004)
9. Dequied, V.: Efficient collusion in optimal auctions. Journal of Economic Theory 136, 302–323 (2007)
10. Deltas, G.: Determining damages from the operation of bidding rings: An analysis of the post-auction knockout sale. Economic Theory 19, 243–269 (2002)
11. Graafland, J.J.: Collusion, reputation damage and intereste in codes of conduct: the case of a Dutch construction company. Business Ethics: A Europesn Review 13, 127–142 (2004)
12. Graham, D.A., Mashall, R.C.: Collusive bidder behavior at a single object second-price and English auctions. Journal of Political Economy 95(6), 1217–1239 (1987)
13. Hu, A., Offerman, T., Onderstal, S.: Fighting collusion in auctions: an experimental investigation. International Journal of Industrial organization 29, 84–96 (2011)
14. Preston Mcafee, R., Mcmillan, J.: Bidding rings. American Economic Review 82, 579–599 (1992)
15. Porter, R.H., Zona, J.D.: Detection of bid rigging in procurement auction. Journal of Political Economy 101, 518–538 (1993)
16. Pesendorfer, M.: A study of collusion in first-price auctions. Review of Economic Studies 67, 381–411 (2000)
17. Sherstyuk, K.: Collusion in private value ascending price auctions. Journal of Economic Behavior & Organization 48, 177–195 (2002)
18. Gou, H., Li, C.: Two Cartel collision mechanism in auction. Journal of Industrial Engineering and Engineering Management 84, 130–133 (2008) (in Chinese)

Traditional Inventory Models for Better Price Competitiveness

Martina Hedvièáková and Alena Pozdílková

University of Hradec Kralove, Czech Republic
{martina.hedvicakova,alena.pozdilkova}@uhk.cz

Abstract. Key factor success in logistics management is cost effectiveness. This article aims to describe and apply a method Economic order quantity (EOQ), which allows managers to make a number of important supply decisions. Managers can use EOQ to determine the quantity of items ordered and how often to order. When used to determine the size of the batch, then it is called a model of economic lot size. For the lot size problem we can consider various special cases, one of which is the case using Monge properties. It can be shown that for a given case are lot-size problems solvable in linear time.

Keywords: Lot-size problem, Economic Order Quantity, cost, optimization, effectiveness, matrix, management.

1 Introduction

For effective inventory management can be used several methods of inventory planning. The paper focuses on the use of methods Economic order quantity that determines the amount of items ordered and how often to order. EOQ principle is based on the comparison of costs related with large inventories (maintenance and storage of supplies - aging, breakage, insurance, inventory management) and costs associated with too little inventory (cost of inventories - buying process and administration, transportation, inventory inspection , price, if dependent on the size of the order). To determine the size of the production batch will be applied to model lot-sizing problem. In the example it is shown that the economic lot-size problem can be solved with linear complexity for the Monge matrices. These matrices are very useful in many practical applications, for example for solving two-sided systems of linear equations [16] or the traveling salesman problem [17], authors also can construct these matrices and formulate many useful theorems [18]. This article shows another important application of these matrices.

2 Economic Order Quantity (EOQ)

Is one of the important techniques used to determine the optimum quantity or number of orders to be placed from the suppliers. The main objective of EOQ is to minimise the cost of ordering the cost of carrying materials, and total

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_61, © Springer International Publishing Switzerland 2014

cost of production. Ordering costs include cost of stationery, salaries to those engaged in receiving and inspecting, general office and administrative expenses of purchase department. Carring costs are incurred on stationery, salaries, rent, materials, handling cost, interest on capital, insurance cost, risk of obsolescence, deterioration and wastage of material and evaporation. [4] The quantity to be ordered should be such that it minimises the carrying and ordering costs. The exact quantity to be ordered at a time so as to achieve this objective is known as EOQ or Re-order Quantity or Economic Lot Size. The EOQ technique can be determined by tabular method, formula method and graphic method [4].

The formula for EOQ can also be used for determining the optimum ordering quantity as given bellow:

$$EOQ = \sqrt{\frac{2AB}{CS}}$$

where A = Annual Consumption in Units, B = Buying Cost per Order, i.e., cost of ordering and receiving the goods per order, C = Cost per Unit, S = Storage and Carrying Cost per annum, i.e., holding of Inventory per year. [4]

The sum up, EOQ is determined keeping in view the ordering costs and carrying costs. With the interaction of these two costs, the economic ordering costs during that period and total cost to order and carry is the lowest as is made clear in Fig. 1. [4]

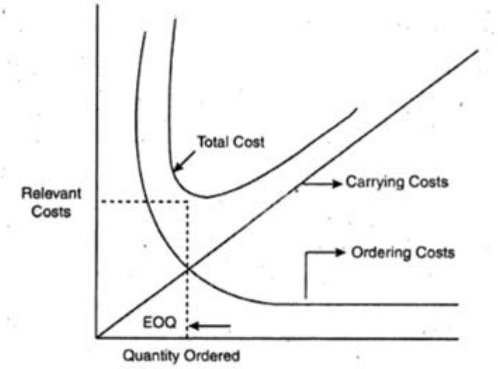


Fig. 1. Economic Order Quantity [4]

Fig. 1 clearly shows the behaviour of the carrying cost, the ordering cost and the sum of these two costs. The carrying cost varies directly with the order size, whereas the ordering cost varies inversely with the order size. [4]

2.1 Illustration Method EOQ – First Example

A company uses a particular material in a factory, which is 10, 000 units per year. The cost per units of material is Rs 15. The cost of placing one order is

Ps 100 and the inventory carrying cost is 25% on average inventory. From this information will be calculate EOQ.

Determination of EOQ

$$EOQ = \sqrt{\frac{2AB}{CS}}$$

$$EOQ = \sqrt{\frac{2 * 10.000 * 100}{15 * 25}} = 730$$

The optimum quantity is 730 units.

2.2 Illustration of Method EOQ – Second Example

The example stores the delicacies from the entire range selected delicatessen salads, which forms the dominant share of total sales. It is distributed in 250 g packages and a 3-week storage period. The salad is expected steady income and sales.

In Graph 1 are recorded orders and sales delicatessen salads to the store. From the graph it can be seen that the highest orders (blue column) in the last quarter of 2012 due to increased Christmas shopping. The lowest income was in the summer months when customers buy alternative products which in summer are not as perishable. Furthermore, the amount of orders was irregular during the year. It was higher due to sales events at discounted prices in April and October.

The sales of salads reacted similarly. The highest sale was in 2012 in the last quarter and during the event to the product (April, October).

The optimal order quantity is:

$$EOQ = \sqrt{\frac{2AB}{CS}}$$

$$EOQ = \frac{15500}{48}$$

The optimal order quantity for this item is 323 pcs salad.

The frequency in days is:

$$f = \frac{365}{48} = 7,6$$

or we can compute the frequency in monhts:

$$f = \frac{12}{48} = 0,25$$

Frequency of order cycle in delicatessen salad is $7,6$ days.

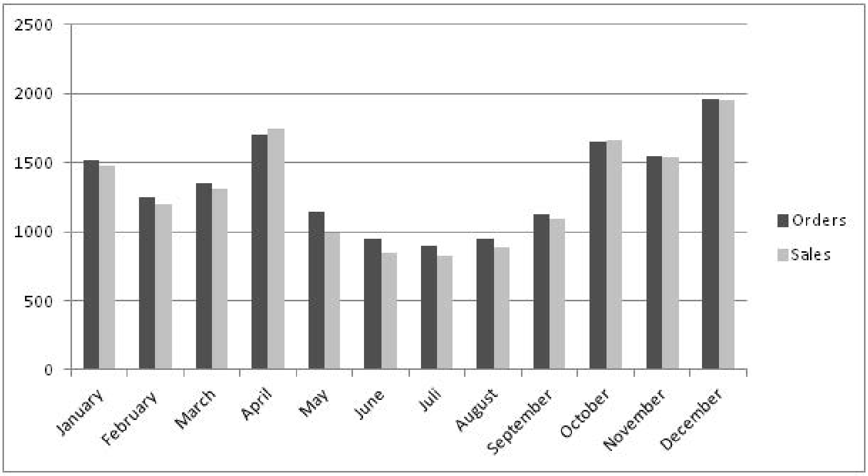


Fig. 2. Order and sale of salads in 2012 (in pieces)

2.3 Computer Solving of EOQ

The Economic Order Quantity models can be also solved by using MS Excel [14]. By this model we can easily compute total costs, average costs and many others. It will be determined the optimal order quantity and total costs in this example.

Variables:

D = Annual Demand Quantity of Product $= 1,000$
$P(c)$ = Purchase Cost per Unit $= 1$
$C(K)$ = Fixed Cost per Order$= 20$
H = Annual Holding Cost per Unit$= 0,25$

EOQ. Economic order quantity $= \sqrt{\frac{2D*K}{h}} = \sqrt{\frac{2*1.000*20}{0.25}} = 400$ units.
Number of orders per year (based on EOQ) $= \frac{1,000}{400} = 2,5$ orders per year

Total Costs. Since the demand will be satisfied with the unit purchase cost any way, it tis discarded from the model. The cost in consideration is reduced to 2 Types: Holding Cost and Ordering Cost. The tradeoff between these costs is optimized at the minimum point of the Total Cost Curve, i.e. EOQ. OEQ is the level of the inventory where ordering cost and carrying cost remains equal. In this section the authors based on the following publications [14], [15].

$$\text{Total Cost} = \text{purchase cost} * \text{production cost} + \text{ordering cost} + \text{holding cost}$$

Purchase Cost: This is the variable cost of goods: purchase unit price ∗ annual demand quantity. This is c ∗ D. Purchase cost = 1 ∗ 1.000 = 1.000.

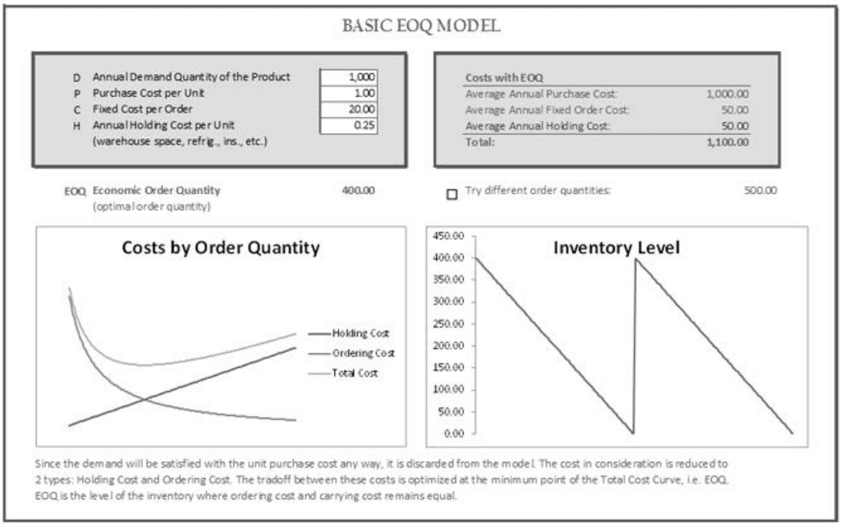
Ordering Cost: This is the cost of placing orders: each order has a fixed cost K, and we need to order D/Q times per year. This is $K * D/Q$. Ordering cost $= 20 * 1000/400 = 50$.

Holding Cost: The average quantity in stock (between fully replenished and empty) is $Q/2$, so this cost is $h * Q/2$. Holding cost $= 0,25 * 400/2 = 50$.

$$TC = c * D + \frac{D * K}{Q} + \frac{h * Q}{2}$$

$$TC = 1000 + 50 + 50 = 1,100$$

Calculating total costs with these values, we get a total inventory cost of \$ 1,100 for the year. This example is illustrated by Fig.3.



BASIC EOQ MODEL

D	Annual Demand Quantity of the Product	1,000
P	Purchase Cost per Unit	1.00
C	Fixed Cost per Order	20.00
H	Annual Holding Cost per Unit (warehouse space, refrig., ins., etc.)	0.25

Costs with EOQ	
Average Annual Purchase Cost:	1,000.00
Average Annual Fixed Order Cost:	50.00
Average Annual Holding Cost:	50.00
Total:	1,100.00

EOQ Economic Order Quantity (optimal order quantity) 400.00

□ Try different order quantities: 500.00

Since the demand will be satisfied with the unit purchase cost any way, it is discarded from the model. The cost in consideration is reduced to 2 types: Holding Cost and Ordering Cost. The tradoff between these costs is optimized at the minimum point of the Total Cost Curve, i.e. EOQ. EOQ is the level of the inventory where ordering cost and carrying cost remains equal.

Fig. 3. Basic EOQ model in MS Excel [14]

3 The Economic Lot-Sizing Problem

A considerable amount of effort has been spent on studying and developing efficient solution procedures for the economic lot-sizing problem. This problem had been solved in 1958, but there is still continuing interest in the problem. The main reason for the continuing interest in this problem is its practical applications. For example, economic lot-sizing is the core problem in aggregate production planning in MRP systems (Nahmias). For an extensive review, see Aggarwal and Park, Bahl at al., Belvaux and Wolsey, Nemhauser and Wolsey, and Wolsey. [1]

The economic lot-sizing problem can be defined as follows. Given the demand, the unit production cost, the unit inventory holding cost for a commodity, the production capacities, and the setup costs for each time period over a finite, discrete-time horizon, find a production schedule that would satisfy demand at minimum cost. [1]

This model assumes a fixed and a variable component of production costs. The fixed cost consists of manpower and materials to start up the machines. To reduce the fixed cost per unit, large lot sizes are desired. On the other hand, for every unit produced there are associated production and inventory holding costs, and the total variable cost (production plus inventory) increases with the number of units produced. Solving the lot-sizing problem means finding a production schedule that would satisfy demand at every period and minimize the total of fixed and variable costs. [1]

The work by Harris in 1913 has been cited as the first study of the economic lot-sizing problem that assumes deterministic demands. This model, known as the Economic Order Quantity (EOQ) model, proposes a production schedule to satisfy the demand for a single commodity with a constant demand rate. Production takes place continuously over time, and the model does not incorporate capacity limits. [1]

A major limitation of the above model is that the demand is continuous over time and has a constant rate. Manne and Wagner and Whitin [3] studied the lot-sizing problem with a finite time horizon consisting of a number of discrete periods, each with its own deterministic and independent demand. [1]

3.1 Formulation

The basic economic problem, which occurs for example in the production systems has been studied and described in [10], [11].

Consider a production that is divided into n periods, and let d_i is the demand in the i-th period. This demand can be met from production in the i-th period or x_i also held that remained in production in the past. Further, let $s(x)$ denotes the cost of producing x units held in the i-th period and let $h_i(y)$ denotes the cost of storage y units of goods from the previous period $(i-1)$ at time i, where $i = 1, \dots N$.

If we further require that the opening and closing stocks are zero, then it is possible to plan production $(x_1, \dots, x_n)$ characterized by the following system of constraints for $i = 1, \dots n$, where y_i is the number of units that are transferred from the i-th period to the next:

$$x_i + y_i - y_{i+1} = d_i$$

$$y_1 = y_{n+1} = 0$$

$$x_1 \geq 0, y_i \geq 0$$

The basis of the economic lot-size problem is to find a feasible production plan that minimizes total costs, which satisfy the following equality:

$$\sum_{i=1}^{n} c_i(x_i) = \sum_{i=2}^{n} h_i(y_i)$$

This problem is generally an NP-hard problems, only in the special case where ci and hi are concave functions can be achieved by using dynamic programming quadratic complexity (see [10]).

Let us denote by $E(j)$ the minimum possible cost of production for the first program to (j - 1)-st period such that y_j stocks, transferred to j-th term is zero. Then there is always a production program with minimal cost such that the demand in each j-th period is either fully covered by the production in that period, or in the stock prior to the (j - 1)-st period.

This gives the following recursive equality:

$$E(j) = \min_{1 \leq i < j} \{E_i + c_i(d_{ij}) + \sum_{q=i+1}^{j-1} h_g(d_{qj})\}$$

where

$$d_{ij} = \sum_{q=i}^{j-1} d_q$$

for $1 \leq i < j \leq n+1$.

Thus, $E(n+1)$ provides two optimal production plan in the n-th period. For the above problem, we consider some special cases. You will be given one that uses Monge properties (will be defined in the next chapter).

Assuming linear storage costs and production costs of fixed type:

$$c_i(x) = c_i^0 + c_i^1(x)$$

Thus, matrix A of type of $n * (n+1)$ such that

$$a_{ij} = E(i) + c_i^0 + c_i^1 d_{ij} + \sum_{q=i+1}^{j-1} h_g^1(d_{qj})$$

for $i < j$ and $a_{ij} = \infty$ in other cases, is a Monge matrix.

From there we get

$$E(j) = \begin{cases} E(j-1), & \text{for } d_{ij} = 0; \\ \min_{1 \leq i < n} a_{ij}, & \text{for } d_{ij} > 0. \end{cases}$$

It can be shown (see [10]) that the economic lot-size problem can be solved with linear complexity for the Monge matrices, which will be defined in the next chapter.

4 Monge Matrices

Monge matrices are a special type of matrices. They have been studied mainly in the max-plus algebra, for its great importance in simplifying some algorithms that are for general matrices difficult to solve (such as the traveling salesman problem). In this chapter these matrices will be studied in max-min algebra, where they have many important properties and also to optimize some problems. Monge matrices were investigated for example in publications [12], [13].

Monge matrices in max-min algebra are special case of matrices, which satisfies the following conditions:

$$a_{ij} \otimes a_{kl} \leq a_{il} \otimes a_{jk} \tag{1}$$

for all $a \in A$, where A is a matrix of type (m, n) and all indexes $i,\ j,\ k,\ l\ \in\ N$, $m \geq 2, n \geq 2$, where i, k are row indexes, for which $i < k$ and j, l are column indexes, for which $j < l$.

This condition is evidently fulfilled for all of those four elements of the matrix, which in the matrix form a rectangle.

Condition 1 may also be formulated in the following equivalent shape:

$$a_{ij} \otimes a_{kl} \leq a_{ij} \otimes a_{kl} \otimes a_{il} \otimes a_{jk} \tag{2}$$

The formulation 2 implies that minimum of every quaternion is considered a_{ij} nebo a_{kl}, that is placed in the quaternion in the upper left or the lower right corner. The condition 2 implies the following theorem.

Theorem 1. *The matrix of type* $(2, 2)$ *is Monge, when the minimum element is placed in the the upper left or lower right corner - the position* $(1, 1)$ *or* $(2, 2)$.

Direct way of verifying whether a given matrix is or is not Monge are very time consuming because it is necessary Monge condition verify for all quaternion of elements. To simplify determination whether a given matrix is Monge we use the following theorem.

Theorem 2. *The matrix is a Monge matrix, if all its submatrices of type* $(2, 2)$, *consisting of two adjacent rows and columns are Monge.*

In Monge matrices we can also define minimal elements.

Theorem 3. *If there is only one minimal element in the Monge matrix of type* (m, n), *where* $m \geq 2, n \geq 2$, *then this element is positioned in the upper left or the lower right corner of the matrix - the position* $(1, 1)$ *or* (m, n).

Theorem 4. *If in the Monge matrix of type* (m, n), *where* $m \geq 2, n \geq 2$, *there are multiple minimum elements, these elements are located around the convex top left and bottom right corner of the matrix. Minimum elements may be positioned in both these areas, or only one of them. Monge matrix apart from the above may include one or multiple columns and one or more lines that contain only minimal elements.*

Proves of the theorems above, many other theorems and global construction of these matrices authors describe in [18].

4.1 Examples of Monge Matrices

In following examples are seen minimal elements and properties resulting from Monge property.

Binary Monge matrix:

$$\begin{pmatrix} 0&0&0&1&0&0&1&1&1 \\ 0&0&1&1&0&0&1&1&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&1&1&1&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0 \end{pmatrix}$$

Monge matrix in a decimal system:

$$\begin{pmatrix} 0&3&8&8&0&9 \\ 0&4&8&7&0&6 \\ 5&5&9&6&0&6 \\ 5&6&4&4&0&3 \\ 9&5&4&1&0&1 \end{pmatrix}$$

5 Conclusion

Nowadays the emphasis is on the highest efficiency investments and cost optimization. One way to reduce costs is to optimize inventory. The article describes the basic method of inventory management and related EOQ lot sizing problems. Cost effectiveness has been a key factor underlying success in logistics management, and usually a vendors price discount offer is an important factor in purchasing considerations. In traditional inventory models, to keep price-competitive, a vendor offers discounts based on the quantity ordered. Buyers attempt to order an item in large quantities in spite of incurring an associated storage cost, so as to dilute the high one time setup ordering cost and take advantage of the discount provided by the vendor. In the case of one item, the traditional quantity discount models have been studied extensively in the literature, for both EOQ type models and lot sizing models [2]. The basic EOQ model and formula has been created to solve the problem of the optimal order quantity via the inventory cost minimization. It is not possible to use the model when demand is stochastic. To analyze this situation the statistical methods are possible to use, however there might be complications if there is no historical data or the trend and behavior of the customers varies. This is the reason for presenting the simple simulation model in MS Excel. It is possible to test various ordering policies and to find the best that is acceptable for the retailer as well as for the supplier and the chain as a whole. It also takes into account various demand fluctuations. For the economical lot sizing problems we can consider various special cases, one of which is Monge matrices. Lot-site problems are NP-hard in general, but with using Monge property it they can be solvable in linear time. Using this methods, managers and management can make a number of significant supply decisions.

Acknowledgements. Authors thank to the Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic for supporting the research by grant "ICT as a Tool of Effectiveness and Growth" in original "ICT jako nastroj efektivity a rustu", project number 8/2013, contract number 2108 and also thanks for non-financial support of this research. This paper was supported by the project No. CZ.1.07/2.2.00/28.0327 "Innovation and Support of Doctoral Study Program" (INDOP), financed from EU and Czech Republic funds.

References

1. Floudas, C.A., Pardalos, P.M. (eds.): Enclyclopiea of production, 2nd edn., vol. XXXIV, 4626 p. 613 illus. Springer (2009) ISBN 978-0-387-74760-6
2. Xu, J., Lu, L.L., Glover, F.: The deterministic multi-item dynamic lot size problem with joint business volume discount. Annals of Operations Research 96(1-4), 317–337 (2000) ISSN: 1572-9338
3. Wagner, H.M., Whitin, T.M.: A dynamic version of the economic lot size model. Management Science 5, 89–96 (1958)
4. Periasamy, P.: Financial Management, 2nd edn., 736 p. Tata McGraw-Hill Education (2009) ISBN: 9780070153264
5. Federgruen, A., Lee, A.: The dynamic lot size model with quantity discount. Naval Research Logistics 37, 707–713 (1990)
6. Bregman, R.L.: An experimental comparison of MRP purchase discount methods. Journal of Operational Research Society 42, 235–245 (1991)
7. Silver, E.A., Peterson, R.: Decision Systems for Inventory Management and Production Planning, 2nd edn. Wiley, New York (1985)
8. Johnson, L.A., Montgomery, D.C.: Operations Research in Production Planning, Scheduling, and Inventory Control. Wiley, New York (1974)
9. Buffa, E.S., Miller, J.G.: Production-Inventory Systems: Planning and Control. Irwin, Homewood (1979)
10. Aggarwal, A., Park, J.: Improved algorithms for economics lot-size problems. Oper. Res. 41, 549–571 (1993)
11. Burkard, R.E., Klinz, B., Rudolf, R.: Perspectives of Monge properties in optimization. Discrete Applied Math. 70, 95–161 (1996)
12. Burkard, R.E.: Monge properties, discrete convexity and applications. European J. Oper. Res. 176, 1–14 (2007)
13. Gavalec, M., Plavka, J.: Structure and dimension of the eigenspace of a concave Monge matrix. Discrete Applied Math. 157, 768–773 (2009)
14. Basic EOQ model, `http://office.microsoft.com/en-us/templates/basic-eoq-model-TC010370172.aspx` (cit. May 20, 2013)
15. Economic order quality, `http://www.answers.com/topic/economic-order-quantity` (cit. May 20, 2013)
16. Krbalek, P., Pozdilkova, A.: Maximal solution for two sided problem in max-min algebra. Kybernetika 46(3), 501–512 (2010)
17. Pozdílkova, A.: Usage of the extremal algebra in solving the traveling salesman problem. In: MME 2012, Slezská univerzita, Karviná (2012)
18. Pozdilkova, A.: Construction of Monge matrices in max-min algebra. In: MME 2011, VE, Janska dolina Slovakia (2011)

Problem of Optimal Route Determining for Linear Systems with Fixed Horizon

Edward Kozłowski

Department of Quantitative Methods,
Lublin University of Technology
Nadbystrzycka 38, 20-618 Lublin, Poland
e.kozlovski@pollub.pl

Abstract. The routing problem of linear system is investigated in this paper. The linear quadratic control problem was reduced to determine the optimal trajectory (way, track, path), which must be tracked by linear system.The general aim of optimal route determining consists of minimization of composite cost function. Moreover, it is compared to the optimal controls for the classical task (LQC) and the task of optimal path determining. To illustrate those controls and track a numerical example is included.

Keywords: optimal route, linear quadratic control, navigation, landmarks.

1 Introduction

The control, navigation, stabilization, costs minimization, identification etc. problems for a different system are widely presented in literature (see e.g. [1]-[5], [7], [11], [13], [17]). In each of these tasks we must control the object to perform the aim. These tasks are connected with optimization. By solving these optimization tasks we determine the control law for system in the explicit form or not. As a result, we can control the object to perform the control aim. Sometimes, in order to achieve the aim the system should be moved after a certain path (trajectory). Thus, the problem arises when there are many guide paths. Which of these trajectories is optimal? In this way, we have the problem of system navigating, where first we must determine the landmarks and next we must lead the system in such a way to mimic these marks (points). The task presented in the article consists of determining the optimal path on which the system achieves the lowest costs. Of course, taking into account other criteria we receive other trajectories (for example, minimizing the entropy of the system during self-learning process, we get a path that gives us the most information about this system).

The tasks of optimal control and route are dual. Namely, if the control law is known then we can specify the path after which the object is to move, and if the optimal trajectory is know then we can control the object to imitated the trajectory.

When selecting the best route and controls the different criteria (shortest path, quickest path, minimal cost etc.) are taken into account. Sometimes the

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_62, © Springer International Publishing Switzerland 2014

Robbins-Monro algorithm is used to solve some technical problems (control, navigation, stabilization, identification, source seeking, see e.g. [3], [8], [12], [13], [15]) . For undetermined number of steps additionally the problem of convergence is verified. In above algorithms the robot (technical system) first is controlled with greater effort and next with little effort. In the present case the problem of movement and control of linear systems is presented, where the main criterion is to minimize the cost. Of course, for a fixed horizon the energy (control) and landmarks are seventy distributed out over time.

The paper presents the problem of determining the optimal trajectory (way, route), after which the controlled system (object, robot) should move. By considering this fact, this paper exploits an idea of dynamic programming. The solution of presented task gives the optimal trajectory (a set of landmarks, statemarks).

The paper is organized as follows. In section 2 the linear quadratic routing problem is formulated and the idea of conversion from control to navigation is outlined. Next, the solution of routing problem is provided in section 3. Section 4 presents an optimal controls and route for the simple linear stochastic system. Additionally the numerical simulation demonstrated for this simple linear system shows that the differences between optimal trajectory and simulated trajectory (path where the system was controlled optimally) are negligible.

2 Linear Quadratic Routing Problem

Sometimes for dynamical systems it is better to determine the optimal route (path) instead controls. Next, the system must be controlled so as to follow a designated path. Therefore, the task is to determine the optimal trajectory after which we want to move the system. In this case the routing means determining the set of points (marks, landmarks), which must be tracked by the system to satisfy the aim, then the system must be controlled. The objective function represents total costs, which are the sum of control costs and costs associated with not hitting the point (target). This total cost is called a composite costs function (CCF). Let $(\Omega, \mathcal{F}, P)$ be a complete probability space. Suppose that $w_1, w_2, ...$ are independent n-dimensional random vectors on this space, with normal $N(0, I_n)$ distribution. We assume that all the above mentioned objects are stochastically independent and an initial state is $\|y_0\| < \infty$.

Let the stochastic linear system be described by a state equation

$$y_{i+1} = Ay_i - Bu_i + C\xi + \sigma w_{i+1} \tag{1}$$

where $i = 0, ..., N-1$, $y_i \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{n\times l}$ and $\sigma \in \mathbb{R}^{n\times n}$. On $(\Omega, \mathcal{F}, P)$ we define a family of sub-σ-fields $\mathbb{Y}_j = \sigma\{y_i : i = 0, 1, ..., j\}$. Below we assume that the parameters of linear system $\xi \in \mathbb{R}^k$ are unknown and has a normal distribution $N(m, Q)$. The matrices $\|B\| < \infty$, $\|C\| < \infty$, $\|\sigma\| < \infty$ where $\|\cdot\|$ denotes a matrix norm as $\|A\| = \max_{\|x\|\leq 1} \|Ax\|$ (the system (1) is Boundary Input Boundary Output stable).

The classical aim of control consists in optimization of performance criterion. For the linear quadratic control problem the aim of control is to minimize the total cost, which is a sum of costs and losses. Then, the task is to find

$$\inf_{u\in U} E\left\{\sum_{i=0}^{N-1} u_i^T R u_i + y_N^T Q y_N\right\} \tag{2}$$

where $\mathbb{Y}_j$-measurable vector $u_j \in \mathbb{R}^l$ is called a control action, and $u = (u_0, u_1, ...)$ an admissible control and the class of admissible controls is denoted by U. At time i the value $u_i^T R u_i$ presents a cost of control and the value $y_\tau^T Q y_\tau$ presents a heredity function as losses (add costs) associated with not hitting to target.

The main aim is to move the system from state y_0 to state origin coordinates $col\,(0, 0, ..., 0)$. The system should be carried out (controlled) the cheapest cost. On the other hand, when we need to determine an optimal route, then the task (2) should be formulated in a slightly different form. Let $det\left(B^T B\right) \neq 0$. When we want to move the system (1) from state y_i to y_{i+1}, $i = 0, 1, ..., N-1$ then the control has a form

$$u_i = -\left(B^T B\right)^{-1} B^T\left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right) \tag{3}$$

Thus, the task (2) may be replaced by

$$\inf_{y\in Y} E\sum_{i=0}^{N-1}\left[\left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right)^T K\left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right) + y_N^T Q y_N\right] \tag{4}$$

where

$$K = B\left(B^T B\right)^{-1} R\left(B^T B\right)^{-1} B^T$$

The main aim is not the solution of problem (2) but only consists in determining the optimal route. Direct solution of task (2) gives us the explicit formula of optimal control, but the solution of task (4) gives the optimal trajectory.

3 Optimal Route Determining

By solving a task (4) we obtain a set of admissible points (marks) $y = (y_0, ..., y_{N-1})$ for which the infimum is attained. The sequence $y_i, i = 0, 1, ..., N$ presents a route (optimal path, trajectory), after which the system (1) should move. Before the recipe (rule) for trajectory planning will be given, we introduce some remarks of filtering conditionally normal sequences.

Remark 1. If the random vector ξ in system (1) has a normal $N(m, \Sigma)$ then, applying the results of the theory of filtering conditionally normal sequences (see e.g. [14], [15]), we have:

1. the conditional distribution $P\left(d\xi\left|Y_j\right.\right)$ is a normal distribution $N\left(m_j, \Sigma_j\right)$

2. the best estimator of the random vector ξ (conditional expectation) $m_j = E\left(\xi \left|Y_j\right.\right)$ and the conditional covariance matrix $\Sigma_j = E\left(\left[\xi - m_j\right]\left[\xi - m_j\right]^T \left|Y_j\right.\right)$ are expressed by formulas

$$m_j = \left(I + \Sigma \sum_{i=0}^{j-1} C^T \left(\sigma\sigma^T\right)^{-1} C\right)^{-1} \left(m + \Sigma \sum_{i=0}^{j-1} C^T \left(\sigma\sigma^T\right)^{-1} \left[y_{j+1} - Ay_j + Bu_j\right]\right)$$

and

$$\Sigma_j = \left(I + \Sigma \sum_{i=0}^{j-1} C^T \left(\sigma\sigma^T\right)^{-1} C\right)^{-1} \Sigma$$

The theorem below presents the method of determining the optimal route, which must be tracked by a system.

Theorem 1. *Let*

$$P_j = A^T \left(K - K^T \left(K + A_{j+1}\right)^{-1} K\right) A \qquad (5)$$

$$L_j = A^T KC - A^T K^T \left(K + A_{j+1}\right)^{-1} \left(KC - L_{j+1}\right) \qquad (6)$$

$$M_j = M_{j+1} + C^T KC + \left(KC - L_{j+1}\right)^T \left(K + A_{j+1}\right)^{-1} \left(KC - L_{j+1}\right) \qquad (7)$$

$$Z_j = Z_{j+1} + tr\left(P_{j+1} H_j\right) + tr\left(\left(M_{j+1} + 2C^T L_{j+1}\right)\left(\Sigma_j - \Sigma_{j+1}\right)\right) \qquad (8)$$

where $P_N = Q$, L_N, M_N *are matrix zero,* $Z_N = 0$ *and*

$$\Sigma_j = E\left(\left(\xi - E\left(\xi \middle| F_j\right)\right)\left(\xi - E\left(\xi \middle| F_j\right)\right)^T \middle| Y_j\right)$$

$$H_j = C\Sigma_j C^T + \sigma\sigma^T$$

If $det\left(K + A_{i+1}\right) \neq 0$ *for* $i = 0, 1, ..., N-1$ *then the optimal state (mark, position) for the time* $j+1$ *based on information available to time* j *is*

$$E\left(y_{j+1} \middle| Y_j\right) = \left(K + P_{j+1}\right)^{-1} \left(KAy_j + \left(KC + L_{j+1}\right) E\left(\xi \left|F_j\right.\right)\right) \qquad (9)$$

and

$$\inf_{y \in Y} E\left(\sum_{i=0}^{N-1} \left[\left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right)^T K \left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right)\right] + y_N^T Q y_N\right) = W_0\left(y_0\right)$$

where

$$W_N\left(y_N\right) = y_N^T Q y_N \qquad (10)$$

$$W_i\left(y_i\right) = y_j^T P_j y_j + 2y_j^T L_j E\left(\xi \left|Y_j\right.\right) + E\left(\xi^T \left|Y_j\right.\right) M_j E\left(\xi \left|Y_j\right.\right) + Z_j \qquad (11)$$

Proof. First we define the Bellmann's function (see e.g. [6], [9], [10]). For the time N the value $W_N\left(y_N\right)$ is given by (10) and for the times $i = 0, 1, 2, ..., N-1$ is defined as

$$W_i\left(y_i\right) = \min_{y_{i+1}} E\left\{\left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right)^T K \left(y_{i+1} - Ay_i - C\xi - \sigma w_{i+1}\right) + W_{i+1}\left(y_{i+1}\right) \middle| Y_i\right\} \qquad (12)$$

for $j = 0, 1, ..., N-1$. From (12) for the time $N-1$ we have

$$\begin{gathered}W_{N-1}\left(y_{N-1}\right) = \min_{y_N} E\left\{ y_N^T \left(K+Q\right) y_N - 2y_N^T K\left(Ay_{N-1} + C\xi + \sigma w_N\right)\right.\\ \left.+2y_{N-1}^T A^T KC\xi + \xi^T C^T KC\xi \right| Y_{N-1}\big\} + y_{N-1}^T A^T KAy_{N-1} + tr\left(\sigma^T K\sigma\right)\\ = \min_{y_N}\left\{E\left(\left.y_N^T\right| Y_{N-1}\right)\left(K+Q\right)E\left(\left.y_N\right|Y_{N-1}\right) + tr\left(\left(K+Q\right)H_{N-1}\right) + tr\left(\sigma^T K\sigma\right)\right.\\ -2E\left(\left.y_N^T\right|Y_{N-1}\right)K\left(Ay_{N-1} + CE\left(\xi\left|Y_{N-1}\right.\right)\right) - 2tr\left(C^T KC\varSigma_{N-1} + \sigma^T K\sigma\right)\\ \left.+2y_{N-1}^T A^T KCE\left(\xi\left|Y_{N-1}\right.\right) + E\left\{\left.\xi^T C^T KC\xi\right|Y_{N-1}\right\} + y_{N-1}^T A^T KAy_{N-1}\right\}\end{gathered}$$

From the properties of condition expectation and matrix properties

$$E\left\{\left.\xi^T C^T KC\xi\right|Y_{N-1}\right\} = E\left(\xi^T\left|Y_{N-1}\right.\right)C^T KCE\left(\xi\left|Y_{N-1}\right.\right) + tr\left(C^T KC\varSigma_{N-1}\right)$$

Hence

$$\begin{gathered}W_{N-1}\left(y_{N-1}\right) = \min_{y_N} E\left(\left.y_N^T\right|Y_{N-1}\right)\left(K+Q\right)E\left(\left.y_N\right|Y_{N-1}\right)\\ -2E\left(\left.y_N^T\right|Y_{N-1}\right)K\left(Ay_{N-1} + CE\left(\xi\left|Y_{N-1}\right.\right)\right) + 2y_{N-1}^T A^T KCE\left(\xi\left|Y_{N-1}\right.\right)\\ +E\left(\xi^T\left|Y_{N-1}\right.\right)C^T KCE\left(\xi\left|Y_{N-1}\right.\right) + y_{N-1}^T A^T KAy_{N-1} + tr\left(QH_{N-1}\right)\end{gathered}$$

The expected optimal state (position, mark) at time N based on information available to time $N-1$ is

$$E\left(\left.y_N\right|Y_{N-1}\right) = \left(K+Q\right)^{-1}K\left(Ay_{N-1} + CE\left(\xi\left|Y_{N-1}\right.\right)\right)$$

and

$$\begin{gathered}W_{N-1}\left(y_{N-1}\right) = y_{N-1}^T A^T\left(K - K^T\left(K+Q\right)^{-1}K\right)Ay_{N-1}\\ +2y_{N-1}^T A^T\left(I - K^T\left(K+Q\right)^{-1}\right)KCE\left(\xi\left|Y_{N-1}\right.\right)\\ +E\left(\xi^T\left|Y_{N-1}\right.\right)C^T\left(K + K^T\left(K+Q\right)^{-1}K\right)E\left(\xi\left|Y_{N-1}\right.\right) + tr\left(QH_{N-1}\right)\\ = y_{N-1}^T P_{N-1}y_{N-1} + 2y_{N-1}^T L_{N-1}E\left(\xi\left|Y_{N-1}\right.\right) + E\left\{\xi^T\left|Y_{N-1}\right.\right\}M_{N-1}E\left\{\left.\xi\right|Y_{N-1}\right\} + Z_{N-1}\end{gathered}$$

We assume, that equation (11) is true for $i+1$. From (11)-(12) and the properties of condition expectation we have

$$\begin{gathered}W_j\left(y_j\right) = \min_{y_{j+1}} E\left\{\left(y_{j+1} - Ay_j - C\xi - \sigma w_{j+1}\right)^T K\left(y_{j+1} - Ay_j - C\xi - \sigma w_{j+1}\right)\right.\\ \left.+y_{j+1}^T P_{j+1}y_{j+1} + 2y_{j+1}^T L_{j+1}E\left(\xi\left|Y_{j+1}\right.\right) + E\left(\xi^T\left|Y_{j+1}\right.\right)M_{j+1}E\left(\left.\xi\right|Y_{j+1}\right) + Z_{j+1}\right|Y_j\right\}\\ = \min_{y_{j+1}}\left\{E\left(\left.y_{j+1}^T\right|Y_j\right)\left(K + P_{j+1}\right)E\left(\left.y_{j+1}\right|Y_j\right) + tr\left(\left(K + P_{j+1}\right)H_j\right) + y_j^T A^T KAy_j\right.\\ +tr\left(\sigma^T K\sigma\right) + 2y_j^T A^T KCE\left(\xi\left|Y_j\right.\right) - 2E\left(\left.y_{j+1}^T\right|Y_j\right)K\left(Ay_j + CE\left(\xi\left|Y_j\right.\right)\right)\\ -2tr\left(C^T KC\varSigma_j + \sigma^T K\sigma\right) + 2E\left(\left.y_{j+1}^T L_{j+1}E\left(\xi\left|Y_{j+1}\right.\right)\right|Y_j\right)\\ \left.+E\left\{\left.\xi^T C^T KC\xi\right|Y_j\right\} + E\left(\left.E\left\{\xi^T\left|Y_{j+1}\right.\right\}M_{j+1}E\left\{\left.\xi\right|Y_{j+1}\right\}\right|Y_j\right) + Z_{j+1}\right\}\end{gathered}$$

and

$$E\left(y_{j+1}^{T} L_{j+1} E\left(\xi \left| Y_{j+1} \right.\right) \middle| Y_{j} \right) = E\left(y_{j+1}^{T} \middle| Y_{j} \right) L_{j+1} E\left(\xi \left| Y_{j} \right.\right) + tr\left(L_{j+1}\left(\Sigma_{j} - \Sigma_{j+1}\right) C^{T} \right)$$

$$E\left\{ \xi^{T} C^{T} K C \xi \middle| Y_{j} \right\} = E\left(\xi^{T} \left| Y_{j} \right. \right) C^{T} K C E\left(\xi \left| Y_{j} \right.\right) + tr\left(C^{T} K C \Sigma_{j} \right)$$

$$E\left(E\left\{ \xi^{T} \middle| Y_{j+1} \right\} M_{N-1} E\left\{ \xi \middle| Y_{j+1} \right\} \middle| Y_{j} \right) = E\left(\xi^{T} \left| Y_{j} \right. \right) M_{j+1}\left(\xi \left| Y_{j} \right.\right) + tr\left(M_{j+1}\left(\Sigma_{j} - \Sigma_{j+1}\right)\right)$$

Hence

$$\begin{aligned} W_{j}\left(y_{j}\right) = \min_{y_{j+1}} \Big\{ & E\left(y_{j+1}^{T} \middle| Y_{j} \right)\left(K + P_{j+1}\right) E\left(y_{j+1} \middle| Y_{j} \right) + y_{j}^{T} A^{T} K A y_{j} \\ & + 2 y_{j}^{T} A^{T} K C E\left(\xi \left| Y_{j} \right.\right) - 2E\left(y_{j+1}^{T} \middle| Y_{j} \right)\left(K A y_{j} + \left(K C - L_{j+1}\right) E\left(\xi \left| Y_{j} \right.\right)\right) \\ & + E\left(\xi^{T} \left| Y_{j} \right. \right)\left(M_{j+1} + C^{T} K C \right) E\left(\xi \left| Y_{j} \right.\right) + Z_{j+1} + tr\left(P_{j+1} H_{j}\right) \\ & + 2tr\left(L_{j+1}\left(\Sigma_{j} - \Sigma_{j+1}\right) C^{T} \right) + tr\left(M_{j+1}\left(\Sigma_{j} - \Sigma_{j+1}\right)\right) \Big\} \end{aligned}$$

Thus, the expected optimal state (position) at time $j+1$ is

$$E\left(y_{j+1} \middle| Y_{j} \right) = \left(K + A_{j+1}\right)^{-1}\left(K A y_{j} + \left(K C - L_{j+1}\right) E\left(\xi \left| Y_{j} \right.\right)\right)$$

and

$$\begin{aligned} W_{j}\left(y_{j}\right) = & \, y_{j}^{T} A^{T} K A y_{j} + E\left(\xi^{T} \left| Y_{j} \right. \right)\left(M_{j+1} + C^{T} K C \right) E\left(\xi \left| Y_{j} \right.\right) + 2 y_{N-1}^{T} A^{T} K C E\left(\xi \left| Y_{j} \right.\right) \\ & - \left(K A y_{j} + \left(K C - L_{j+1}\right) E\left(\xi \left| Y_{j} \right.\right)\right)^{T}\left(K + A_{j+1}\right)^{-1}\left(K A y_{j} + \left(K C - L_{j+1}\right) E\left(\xi \left| Y_{j} \right.\right)\right) \\ & + Z_{j+1} + tr\left(P_{j+1} H_{j}\right) + tr\left(\left(M_{j+1} + 2 C^{T} L_{j+1} \right)\left(\Sigma_{j} - \Sigma_{j+1}\right)\right) \\ = & \, y_{j}^{T} A^{T}\left(K - K^{T}\left(K + A_{j+1}\right)^{-1} K \right) A y_{j} + tr\left(\left(M_{j+1} + 2 C^{T} L_{j+1} \right)\left(\Sigma_{j} - \Sigma_{j+1}\right)\right) \\ & + 2 y_{j}^{T}\left(A^{T} K C - A^{T} K^{T}\left(K + A_{j+1}\right)^{-1}\left(K C - L_{j+1}\right)\right) E\left(\xi \left| Y_{j} \right.\right) + Z_{j+1} + tr\left(P_{j+1} H_{j}\right) \\ & + E\left(\xi^{T} \left| Y_{j} \right. \right)\left(M_{j+1} + C^{T} K C + \left(K C - L_{j+1}\right)^{T}\left(K + A_{j+1}\right)^{-1}\left(K C - L_{j+1}\right)\right) E\left(\xi \left| Y_{j} \right.\right) \\ = & \, y_{j}^{T} P_{j} y_{j} + 2 y_{j}^{T} L_{j} E\left(\xi \left| Y_{j} \right.\right) + E\left(\xi^{T} \left| Y_{j} \right. \right) M_{j} E\left(\xi \left| Y_{j} \right.\right) + Z_{j} \end{aligned}$$

what finish the proof.

Remark 2. The equation (9) gives the formula (recipe, rule) how to determine the optimal route (state- or land- marks) for time $j+1$ if the system (1) to time j traveled the way (path, track) $y_0, ..., y_j$.

Remark 3. Of course, to determine the path on which the object should move, we can act in another way. First we solve the classical linear quadratic control problem (2) and obtain the control laws. Next we simulate the possible trajectories of linear system (1) using the optimal controls. Finally, averaging the possible paths we obtain the trajectory, which must be tracked by the system. The obtained trajectory can not be optimal.

4 Optimal Controls and Route for Linear Quadratic Problem

Let us consider a linear system with state equation

$$y_{i+1} = y_i - Bu_i + \sigma w_{i+1} \tag{13}$$

The optimal control of linear system (13) for the task (2) contains in follow

Lemma 1. *If det* $\left(R_i + B^T G_{i+1} B\right) \neq 0$ *for* $i = 0, 1, ..., N-1$ *where*

$$G_i = G_{i+1} - G_{i+1}^T B \left[R_i + B^T G_{i+1} B\right]^{-1} B^T G_{i+1} \text{ and } G_N = Q \tag{14}$$

then the optimal control is

$$u_i^* = \left[R_i + B^T G_{i+1} B\right]^{-1} B^T G_{i+1} y_i \tag{15}$$

and

$$\inf_{u \in U} E \left\{ \sum_{i=0}^{N-1} u_i^T R_i u_i + y_N^T Q_N y_N \right\} = y_0^T G_0 y_0 + \sum_{j=1}^{N} tr \left(\sigma^T G_j \sigma\right)$$

Proof of this lemma we can find in [12], [13].

Remark 4. The optimal route (trajectory, set of landmarks) for the system (13) is

$$E\left(y_{j+1} | F_j\right) = \left(K + A_{j+1}\right)^{-1} K y_j \tag{16}$$

$j = 0, 1, ...N-1$ where A_j is defined as (5) When we want to plane a trajectory (route, path) at time $t = 0$ then we must determine optimal route conditioned on $\sigma-$field Y_0

$$E\left(y_{j+1} | Y_0\right) = \left(K + A_{j+1}\right)^{-1} K E\left(y_j | Y_0\right)$$

or in dynamical form

$$E\left(y_j | Y_0\right) = \left(\left(K + A_{j+1}\right)^{-1} K\right)^j y_0$$

Remark 5. When the optimal route for the linear system is known and calculated as (16) then from (13) the expected control conditioned on $\sigma-$field Y_j is

$$\begin{aligned} E\left(u_j | Y_j\right) &= -\left(B^T B\right)^{-1} B^T \left(E\left(y_{j+1} | Y_j\right) - y_j\right) \\ &= \left(B^T B\right)^{-1} B^T \left(K + A_{j+1}\right)^{-1} (I - K) y_j \end{aligned} \tag{17}$$

Table 1. The trace planning and landmarks, simulation of states, optimal controls for linear system

j	$E\,(y_j\|\,Y_0)$	y_j	$E\,(y_j\|\,Y_{j-1})$	u_j
0	$(30, 25)$	$(30, 25)$	——	$(1.797; -0.076)$
1	$(23.72; 25.85)$	$(26.51; 25.90)$	$(23.72; 25.85)$	$(1.755; 0.139)$
2	$(18.38; 25.11)$	$(22.75; 24.35)$	$(20.90; 25.68)$	$(1.648; 0.301)$
3	$(13.94; 23.20)$	$(19.34; 22.56)$	$(18.09; 23.32)$	$(1.543; 0.449)$
4	$(10.30; 20.47)$	$(15.96; 20.04)$	$(15.56; 20.61)$	$(1.408; 0.574)$
5	$(7.335; 17.19)$	$(12.96; 17.17)$	$(12.86; 17.15)$	$(1.290; 0.677)$
6	$(4.914; 13.544)$	$(9.934; 14.14)$	$(10.36; 13.38)$	$(1.148; 0.772)$
7	$(2.886; 9.712)$	$(7.246; 10.44)$	$(7.574; 9.525)$	$(1.041; 0.809)$
8	$(1.098; 5.831)$	$(4.773; 6.804)$	$(4.925; 5.110)$	$(0.975; 0.882)$
9	$(-0.600; 2.033)$	$(2.594; 3.415)$	$(2.025; 0.373)$	$(1.065; 0.819)$
10	$(-0.024; 0.040)$	$(-0.038; 0.126)$	$(0.054; 0.030)$	——

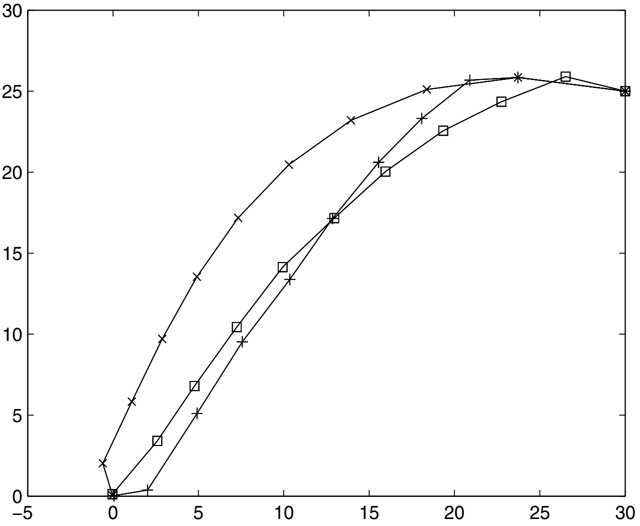


Fig. 1. States simulation, planned trajectory and landmarks for linear system

Example 1. Let us determine the optimal route and controls for a linear system with state equation (13) where the initial state y_0 is $(30; 25)$ and the fixed horizon $N = 10$. This system must be moved to origin coordinates. Let us assume that

$$Q = \begin{bmatrix} 12 & 2 \\ 2 & 8 \end{bmatrix}, \quad R = \begin{bmatrix} 1.2 & 0.2 \\ 0.1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0.5 \\ 0.1 & 4 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 0.2 & -0.03 \\ 0.02 & 0.5 \end{bmatrix}$$

For this case the route planning $E\,(y_j|\,F_0)$, simulated states y_j and landmarks $E\,(y_{j+1}|\,Y_j)$ (expected optimal states conditioned on information to time j), optimal controls u_j are given in the table 1. The third column of the table 1 presents a possible trajectory y_j for the system (13), where the controls at times $j = 0, 1, ..., N - 1$ are optimal.

The figure 1 shows the next curves: the curve with "square" marker (□) presents the simulation states, the curve with "plus sign" marker (+) - landmarks (expected optimal states conditioned on information it time j), the curve with "cross" marker (x) - planned route. We see that the planned route, landmarks, simulated states are very close to each other so that the differences between the landmarks and simulated states are negligible.

5 Conclusion

In this article, the optimal route problem of stochastic discrete-time linear system with quadratic objective function for fixed horizon was presented. The described problem is an idea of conversion from control to navigation of linear system. To determine optimal trajectory the algorithm of dynamic programming was used. As a result we have a set of landmarks. To perform the goal the system (robot, object) must track the optimal path (trajectory).

The extension of described problem can be used, for example, to the source seeking problem, route determining with reach information, planing navigation, perfect tracking etc.

References

1. Aoki, M.: Optimization of Stochastic Systems. Academic Press (1967)
2. Abouzaid, B., Achhab, M.E., Wertz, V.: Feedback stabilization of infinite-dimensional linear systems with constraints on control and its rate. European Journal of Control 17(2), 183–190 (2011)
3. Azuma, S., Sakar, M.S., Pappas, G.J.: Stochastic Source Seeking by Mobile Robots. IEEE Transactions on Automatic Control 57(9), 2308–2321 (2012)
4. Banek, T., Kozłowski, E.: Adaptive control of system entropy. Control and Cybernetics 35(2), 279–289 (2006)
5. Banek, T., Kozłowski, E.: Active and passive learning in control processes application of the entropy concept. Systems Sciences 31(2), 29–44 (2005)
6. Bellman, R.: Adaptive Control Processes. Princeton (1961)
7. Bubnicki, Z.: General approach to stability and stabilization for a class of uncertain discrete non-linear systems. International Journal of Control 73(14), 1298–1306 (2000)
8. Chena, Y., Edgarb, T., Manousiouthakisa, V.: On infinite-time nonlinear quadratic optimal control. Systems and Control Letters 51(3-4), 259–268 (2004)
9. Fleming, W.H., Rishel, R.: Deterministic and Stochatic Optimal Control. Springer, Berlin (1975)
10. Harris, L., Rishel, R.: An algorithm for a solution of a stochastic adaptive linear quadratic optimal control problem. IEEE Transactions on Automatic Control 31(12), 1165–1170 (1986)
11. Kozin, F.: Stability of stochastic dynamical systems. Lecture Notes in Mathematics, vol. 294, pp. 186–229 (1972)

12. Kozłowski, E.: The linear quadratic stochastic optimal control problem with random horizon at finite number of events intependent of state system. Systems Science 36(3), 5–11 (2010)
13. Kozłowski, E.: Identyfication of linear system in random time. International Journal of Computer and Information Technology 1(2), 103–108 (2012)
14. Liptser, R.S., Shiryaev, A.N.: Statistics of Stochastic Processes. Springer, New York (1978)
15. Manzie, C., Krstic, M.: Extremum seeking with stochastic perturbation. IEEE Transactions on Automatic Control 54(3), 580–585 (2009)
16. Saridis, G.N.: Stochastic processes, estimation and control: the entropy approach. John Wiley and Sons (1995)
17. Zabczyk, J.: Chance and decision. Scuola Normale Superiore, Pisa (1996)

Towards Service Science: Recent Developments and Applications

Katarzyna Cieślińska[1], Jolanta Mizera-Pietraszko[1], and Abdulhakim F. Zantuti[2]

[1] Institute of Computer Science, Wroclaw University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370, Wroclaw, Poland
{katarzyna.cieslinska,jolanta.mizera-pietraszko}@pwr.wroc.pl
[2] Faculty of Engineering, Zaytona University,
Wybrzeże Wyspiańskiego 27, 50-370, Tripoli, Libya
abdulhakimzent@hotmail.com

Abstract. The study reports some of the most significant advances in the field of Service Science. In particular, discussed are: Service Composition, Knowledge Engineering and Resource Allocation. We focus on the following applications of service-based systems: eHealth, eLearning and Social Networks.

Keywords: Service Science, Service Systems, SOA, e-Health, eLearning systems, Social Network.

1 Introduction

Current economic environment is radically changing. Evolution in global economic trends, for instance, demographic shift, rapid development of web-based technologies, lead us to understand the advances in terms of how to design, improve and scale service systems for business and social purposes [1].

Services are the basis of today's economy. Therefore, more efficient management of services should be a priority. Companies report an increase in their profits from services rather than production. Thus, Daniel Bell's [2] predictions of 1973 that over the next decades knowledge-based services would outperform manufacturing as a source of employment, came true by the year 2000. The way of stimulating service popularity was to create Service Science discipline, that addresses the problem by integrating science, technology and business to improve productivity, quality and innovation in services [3].

1.1 Service Science

Service Scince integrates organizations' and human understanding with business and technological knowledge [4]. Service Science deals with the analysis and development of complex Service Systems, in which various kinds of software components, people, etc. provide services to others. These systems constitute a dynamic configuration of people, technology, organizations, and shared information.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_63, © Springer International Publishing Switzerland 2014

According to [5], services are becoming the key factor in stimulating the most industrialized economies. Service Systems incorporate people into the process of developing new technologies. The growth of Service Science causes market to change the goods-dominant logic [6] concept to service-dominant logic (S-D logic) [7], [8], [9]. Additionally, according to [10], [11], [12] it applies also to Service Science, Management, Engineering and Design (SSMED). These concepts provide all prerequisites to build a theory of Service Systems.

Service-oriented Computing (SOC) [13] is a paradigm of system development. An application assumes the autonomy and heterogeneity of the components that make up the system. According to this paradigm, one can create different types of software architectures, in particular Service Oriented Architecture (SOA). Services, their descriptions and operations (publication, discovery, selection, and binding) that produce, or utilize such descriptions constitute the foundation of SOA [18], [19].

1.2 Definitions of Service Systems

The idea of a new scientific discipline called Service Science has its genesis in a phone conversation of 2004 between Jim Spohrer(IBM® Research Service department) and Henry Chesbrough (professor of business and innovation at the University of California at Berkeley). The reason was simple: lack of candidates who had the right mix of knowledge including computer science, engineering, management and social science. Over the years, the definitions of Service Systems have evolved. The earliest definitions were created in 2007.

“**Service Systems** represent value-co-creation configurations of people, technology, value propositions connecting internal and external service systems, and shared information (e.g., language, laws, measures, and methods)” [5].

The same year, Qui, Fang, Shen and Yu [14], present their definition.

“**Service Systems** can simply be a software application, or a business unit with an organization from a project team, a business department, a global division; it can be a firm, institusion, government agency, town, city or nation; it can also be a composition of numerous collaborativelu connected service systems within, and/or across organizations.”

In 2008 and 2009, presented were at least three new definitions of service system [15], [16]. The most interesting was by Vargo, Maglio and Akaka [17].

“Every **Service Systems** is both a provider and client of service that is connected by value propositions in value chain, value networks, or value-creating systems.”

A development of economic trends and new technologies have changed the way of doing things. Service Science is evolving what makes the definition more and more complete and precise. Nowadays, services are more custom-oriented and they include more knowledge. Barile and Polese [1] present summary of most relevant Service Systems definitions, displayed in various networks.

1.3 Challenges for Service Science

Service Science is at its primary stage, therefore there are still many doubts about definition and basic concepts.-

In [18], eight challenges for Service Science are stated as a potentially valuable direction in the development and research. Alter [18] informs that given these challenges may be found controversial. Some of them are under exploration (Challenge#8: Maintain analytical rigor without losing the spirit of service), some are just extension of previous ones (Challenge#4: Replace "the customer" with clear distinctions between various customer groups and other stakeholders whose different interests may be in conflict. Challenge#5: Highlight customer and provider responsibilities for value creation). Some of them seem to have become standard, like the definition of Service and Service System(Challenge#1: Use a broadly applicable definition of service, Challenge#2: Use a broadly applicable definition of Service System).

Service Science is related to disciplines such as Knowledge Engineering, Service Composition or Resource Allocation. The paradigms described in Section 1.1 help to understand the challenges posed on Service Science. Scientists build applications using avaible knowledge and technology. They try to connect provider with the client in value systems, presenting much more customized solutions.

The rest of this paper is organized as follows. In Section 2, we review recent developments in Service Science according to the following categories: Service Composition, Knowledge Engineering and Resource Allocation. In Section 3, we highlight that the recent developments are applicable in many domains of software engineering. We divide them into three categories: eHealth systems, eLearning systems, and Social Networks. Summary of the paper is presented in Section 4.

2 Recent Developments

2.1 Service Composition

In general, service composition process consists of two steps: the first one constitutes the required functionalities and their interactions–i.e. control and data flow – they are identified. The second one - for a set of the functionalities appropriate candidate services are discovered from the repository and then selected in an optimization task, resulting in a composite service execution plan that specifies a required composite service. Service composition is a simple way for delivering new functionalities and adapting the existing ones to the changing user requirements. The characterisitics of the web services allow this process to be automated, so that new functionalities can be added depending on the user requirements. This makes Service Oriented Architecture combined with automated service contents, which is an expected solution to a problem of dynamically changing user requirements.

In the literature there are many approaches to service composition problem [19], [20], [21]. Stelmach [22] presents automated negotiation of communication protocols in a composition of data stream processing services and he introduces a planning-based approach. Fraś [23] introduces Smart Service Workbench (SSW) - an integrated tool to support business processes in IT. It consists of modules which cover all the functionalities of the the client-SWW model. Modules are responsible for e.g.: requirements analysis, services choice or service composition.

Other approaches concentrate on a more complex service composition [24]. Grzech [25] describes a translation of service level agreement (SLA) into structure of complex services. In [26] a model of complex services composed of atomic services available in different versions and offered in a heterogeneous environment is shown. The Authors hightlight the aim of the model as a framework to analyze access limitations, services costs, security, and resource constraints, i.e. In [27], a novel approach to the problem of optimization of complex service execution plan is introduced by applying an algorithm to the multidimensional knapsack problem (MKP) solution.

However, the problems of semantic analysis of user requirements, service discovery and selection of services against non-functional requirements still arised. The biggest disadvantage of the presented solutions is that in general, they satisfy either functional [28], or non-functional [29], [30] requirements. In [31], the service composition problem in the Internet of Things paradigm is discussed. An algorithm for composite service plan optimization selects the services according to their functional and non-functional requirements [32], [33]. Presented innovations use a fitness function [34], [21]. The Authors present improvements to the algorithm in relation to the uncertainty aspect of service composition. In [35], the problem of ICT service mapping in service composition process is presented. The Authors formulate the problem of ICT service mapping in the process of service composition and propose a solution. The presented solution applies a concept of decision tables [36], [37] as an ICT service mapping tool.

Also, a framework-based approach in the SOA field was presented [38]. In this work, the authors propose a software framework that incorporates various composition approaches and use different knowledge repositories such as ontologies, social networks, or rule engines.

2.2 Knowledge Engineering

Knowledge Engineering (KE) refers to all technical, scientific and social aspects of building, maintaining and operating knowledge-based systems. KE is an engineering discipline that integrates knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise [39]. It is a multidisciplinary field, bringing in both concepts and methods from several computer science domains such as artificial intelligence, databases, expert systems and decision support systems.

Various techniques from the Knowledge Engineering filed are applied in the Service Science domain. The solutions mainly aim to improve the process of service selection [40], [41] support decision making in the systems based on SOA paradigm [42], or to ameliorate the process of resource allocation [43].

The key issue of knowledge engineering is to propose the architecture of delivering data mining solutions to the systems. In work [44] authors present the service-oriented paradigm of designing the data processing techniques as Web services. Each of the services communicate using standarized interfaces provided by the SOAP protocol. The solutions are widely accessible as connected by ESB.

Huang [45] takes into consideration diverse types of knowledge related to remote sensing image understanding. In their article, discussed is knowledge representation (KR) architecture as classified into six types. Agents for KR task are employed to bridge the gap between low-level image processing methods and high-level semantic descriptions. In addition, the authors of [46] present KR and reasoning using fuzzy Petri nets (FPN). Because it still has many deficiencies, they present a knowledge acquisition and representation approach using fuzzy evidential reasoning approach and dynamic adaptive FPNs. The proposed approach can not only capture experts' various experience well, it extends the KR power, and reason the rule-based knowledge more accurately.

2.3 Resource Allocation

The field of Service Science attracted the researchers attention and the industry in the area of Resource Allocation (RA). RA is the process of determining the best way to use available assets or resources in the completion of a given project. Companies attempt to allocate resources in a manner that helps to minimize costs while maximizing profits at the same time. Recently developed information and communication technologies (ICT) empower entrepreneurs to inbuild monolitic architectures into the distributed ones. Due to the fact, service-oriented architecture (SOA) becomes important paradigm in designing service-oriented systems (SoS). Resources given to services need to be efficiently managed. To ensure high quality of service (QoS), a Resource Allocation dilemma should be considered [47], [48], [49].

Rygielski and Tomczak [41] state the problem and present an algorithm for detection in streams of requests. Dissimilarity measure between two probability distributions is presented with emphasis on long-lasting changes. To estimate the reference and model distributions the sliding window technique was presented.

Tomczak [50], outlines an on-line change detection algorithm for resource allocation, based on the dissimilarity measure framework [51], [52] between two parameterized probability distributions (pds). Author takes advantage of the fact that streams of requeasts in service- oriented systems can be modeled by non-homogenous Poisson processes. In this work, three dissimilarity measures are considered i.e. Bhattachryya distance measure, Kullback-Leibler divergence and an absolute mean difference, in relation to analytical formulae for Poisson distributions. Additionally, resource allocation is initialized, and in case of detected change, the resources are re-allocated due to a given optimization algorithm, e.g., interior-point algorithm [53]. At the end, simulation study is presented.

Song, Sun and Shi [54] propose a set of on-demand resource allocation algorithms based on the dynamic resource allocation mechanism and model. Presented algorithms ensure performance of critical applications, identified by the data center manager. They argue that the existing techniques relying on turning on, or off the servers by a virtual machine (VM), is not enough to solve the problem of Resource Allocation.

Świątek [55] shows a problem of mobility and resource management in heterogeneous wireless networks. He demonstrates that based on the knowledge gathered about the clients' activity, it is possible to predict their future interaction with the system which results in improving the overall quality of the services delivered as well as the network resources utilization. Whereas Grzech [56] discusses multistage processing of connections in connection-switched networks. On introducing the general idea of multistage traffic processing, he argues about a connection classification task in a two-level processing schedule.

3 Applications

Many techniques from the service composition, knowledge engineering or resource allocation have been implemented in the Service Science domain. Here we classify them into three groups.

3.1 eHealth Systems

Recently, paper-based medical systems have been replaced by eHealth systems due to their convenience and accuracy. Also, since the medical data can be stored on any kind of digital devices, people can easily access medical services at any time and place.

The new wireless technology has offered many advantages in the conventional healthcare system like e.g. a special body patch which transmits a patient's health data wirelessly to a GP's practice. The device allows the patient health to be monitored 24 hours a day.

Recently, Body Area Networks (BANs) (or Body Sensor Networks (BSNs)) propose a promising approach to help improve health care system [57]. Initial applications of BANs appeared in the healthcare domain, in particular for continuous monitoring and logging vital parameters of patients suffering from chronic diseases such as diabetes, asthma and heart attacks. However, the design of their eHealth system comes with emerged challenges. One of them is how to ensure security and privacy of the patients' Personal Health Information (PHI) [58], [59]. Apart from security and privacy aspects, development of analytical methods for such systems is analysed. Furthermore, Grzech [60] introduces a problem of e-health service quality management, especially when communicaton conditions are changing.

PAAS

Linke Guo [61] proposes a framework called PAAS (Privacy-Preserving Attribute-Based Authentication System), which identifies users in eHealth systems while preserving their privacy. The Authors propose some authentication strategies between the patients themselves, or between the patients and doctors. Based on the data security and an efficiency analysis, in terms of privacy preservation, the framework turned out to be better than the working eHealth systems.

SAGE

SAGE [62] is a Strong Privacy-Preserving Scheme Against Global Eavesdropping for eHealth Systems. It can preserve not only content-oriented privacy, but also contextual privacy against a strong global adversary. The basic idea of SAGE is quite simple: when the patient information database (PIDB) receives the PHIs from a patient, it transmits them to all the physicians. Then, the potential physicians get the signal from their patients. SAGE can achieve unconditional receiver anonymity. As a consequence, the patient privacy is guaranteed.

In the proposed SAGE, the content-oriented privacy can be guaranteed by the secure symmetric encryption algorithm. To gain that, an adversary always has ways to link the patient to a target physician (RD(PA, PH)--› 0), the trick is to ensure the link between the PIDB and the Physician (PH) (RD(PIDB, PH) ---› 0) by a DP's (programmable demon program) broadcasting which reports (PIR) all the physicians. Privacy link between PIDB and PH is formally defined by a game between a challenger and an adversary.

There may be two kinds of attacks which affect SAGE performance. The Authors state how it is processed. The provoked attack doesn't affect the SAGE performance because the DP controls the validity of the timestamp. If the timestamp is incorrect, it will be discarded. Attack can be also prevented, since the message authentications relies on the static shared key. In addition, digital signature techniques have been integrated in the SAGE.

SmartFit

SmartFit is a system that adopts new technologies of pervasive computing. It was designed to support endurance and technical training of athletes [63]. The data is transmitted between the users of the system (i.e. the athletes and trainers) with predefined quality level, irrespective of their location.

In [64] the general architecture of SmartFit is presented. The Authors distinguish three main functionalities of SmartFit and support them by examples.

- planning volume of endurance training.
- endurance training and monitoring.
- support technical training.

Atomic services are used to provide functionality supporting skill assessment and improvement of the elementary tennis strokes such as serve, backhand and forehand. Based on initial skill level recommendations a support in planning future technical tennis training is proposed. The application allows to support feedback training. The big advantage is that physiological and kinematic data from sensors placed on athlet's body is visible in a real–time. It helps the two parties both to compare the results with the previous ones and with other athlets.

The Authors investigate the process of technical tennis training in detail. In the presented user-case, the scenario for SmartFit athlete's BAN consists of EMG, gyroscopes and accelerometers wireless units. The physiological and kinematic data is transferred to the server. In the skill assessment case, the signals from gyroscopes and accelerometers are processed in order to build a relationship between wrist flexion, upper arm rotation and racquet speed. Obtained results can be compared with the results captured from the reference data of a high-level tennis player. Based on this data it is possible to build a personalised model for the tennis player. It may also be used to make recommendations for future technical training.

eDiab

In [65] introduced is a new method for decision rules extraction called Graph-based Rules Inducer (GRI) to support the medical interview in the diabetes treatment. The Authors present a method for knowledge extraction in a form of decision rules to support anamnesis. This method is implemented in a system called **eDiab** [66].

This work contributes to proposing an algorithm for rules induction that enables a physician to conduct personalized medical interview. Two issues are considered by the authors: First, knowledge representation should be chosen. In this work, decision rules are applied. Second, the context is unknown and non-stationary (evolving in time). GRI by authors is decribed as follows: “In the first step, the graph determining the search space is obtained. In order to create only the paths that are coupled with a proper rule, that is, the sign of the path, the algorithm runs backward, that is, from the final vertex to the first layer. Despite the fact that the number of paths in both sets might grow exponentially with respect to the number of features, in many practical cases the number of paths seems to be reasonable. Nevertheless, formulating a rule-based model in problems with many inputs might be intractable, therefore some heuristics should be proposed”.

Two experimental studies are presented in the article. The findings are discussed. The first one involves 13 other methods and a benchmark data set Electricity [67]. The second one is the application in the diabetes treatment and involves five other methods and a data set collected by Michael Kahn, M.D., Ph.D., published in the UCI Machine Learning Repository [68]. Generally, GRI is very promising and should be developed.

3.2 eLearning Systems

Intelligent E-learning systems attract attention because they relate personalized learning to the particular characteristics of the users. Personal learning styles need to be taken into account while planning an elearning process.

In recent years, many personalized elearning systems have been developed [69]. Each of them emphasizes different criteria. In Insprire [85], knowledge level of the learner is taken into account at the stage of planning the lesson. The systems use links to support the learners’ navigation. INSPIRE offers computer-adaptive testing based on the item response. Educe [86] analyzes the time of learning. Learning units comprise different media types such as text, image, audio and multimedia. EDUCE uses the calculated probabilities to chose one out of the four types of material with the highest probability. In [70] classification into groups is discussed. For each group determined is learning scenario. After a few lessons Students sit a test in their individual capacity. If the learning process is acceptable, the student continues learning, otherwise the system modifies the scenario.

Student Courses Recommendation

The formulation of the recommendation problem was first stated in [71], [72], [73] and the problem has been studied extensively since then on. According to [74], Recommender Systems (RS) are classified into three categories based on how the recommendations are made: Content-base recommendations, Collaborative recommendatrions and Hybrid approaches. The authors of [75] present a survey of the RS incorporating the contextual information into the recommendation process to support multicriteria ratings, or provision in a more flexible and less intrusive recommendation process. Other research also includes explainability [76], scalability [77], [78] and privacy [79] issues.

The main goal of the recommendation systems (RS) is to deliver customized information to the users of increasing web-based systems population. RS are used to solve problems coming from different domains, such as: web recommender, personalized newspaper, sharing news, movie recommender, document recommender, e-commerce, Travel and Store recommender, e-mail filtering, music recommender or music list recommender [80].

In its most common interpretation, the recommendation problem is reduced to the problem of estimating ratings for the items that have not been seen by a user. Intuitively, this estimation is usually based on the user's rating some other items based on some other information that is described below. Once we can estimate ratings for these items, we can recommend the item with the highest rating to the user.

In the section, we present two works discussing recommendation systems from the student courses at the University Information System EdukacjaCL using respectively Ant Colony Optimization (ACO) application; Markov Chain Model and Bayesian Inference.

Ant Colony Optimization Application

ACO [80] is a natural computer algorithm that simulates the behavior of living ants. ACO is applied as an information filtering method in RS. It enables us to introduce a new type of hybrid method called integrated HA. In an experiment, the grade determines the e-lecture with final exam to recommend for the user. The Authors applied three algorithms based on ACO:

1.With maximal probability.
2.With one ant and random walking using edges' probabilities.
3.With *K* ants and random walking using edges' probabilities.

ACO methods were compared with filtering methods:

1. Randomrecommender
2. Content-Basedrecommender (CFB)
3. Collaborative Filtering using demographic information rocommender (CF1)
4. Collaborative Filtering using Euclidean metric recommender (CF2)
5. Collaboratice Filtering recommender (CF3)

To compare their methods the Authors applied the following measures: *Mean Absolute Error* (MAE), *Normalized Mean Absolute Error* (NMAE), *Prediction Accuracy* (PA), *Mean Squared Error* (MSE), *Root MeanSquared Error* (RMSE), *Standard Error Variance* (SEV), *Classification Accuracy* (classAcc) that measure the number of predictions and compare them with the grades in the test set.

All of the presented measures show that the ACO method is more stable than the others. Mean Absolute Error (MAE) measure, which shows difference between real value and the predicted one for the ACO-based method, is around 0.4. That means that recommendation lowers the real value by half of the grade.

Markov Chain Model and Bayesian Inference

The following article focuses on an approach of a Markov model as a knowledge representation, and Bayesian inference as a reasoning method [81]. The Authors present the user's choice of a new decision influenced by the former one. This causes that behavior can be described by the Markov chain model. However, the main concern is the recommendation task. Where the user's behavior is modeled on Markov chains, the personalization problem can be stated as a classification task. The new user makes the best decision with the minimal risk. *The Bayesian algorithm as an optimal decision making algorithm is proposed.* Due to non-stationarity in a real-life situation, an adaptive estimation method is presented. The Authors apply incremental learning algorithm [82] to adopt the changes. At the end, an experimental study is presented.

In the experiment, the state in the Markov chain is associated with the language and its level: *English A1, English A2, English B1, English B2, English C, German A1, German A2, German B1, German B2, German C, nothing*. Moreover, it is assumed that each student is described by a following vector: number of all courses, number of classes, number of laboratories, number of lectures, mean of grades, variance of grades, mean of grades from lectures, mean of grades from exercises, mean of grades from laboratories, number of fails (grade F), number of excellents (grade A). Probably, carrying out the experiment for 10 semesters would demonstrate that applying Markov chain model is even more appropriate.

OnLine Lab

The application **Online Lab** [63] is a distributed, service-based computer laboratory benefiting from the IPv6 QoS architecture, which is used to distribute computational tasks while maintaining the quality of service and the user experience. Online Lab (OL) implements an architecture consisting of user interface (OL-UI), core server (OL-CORE), services and computational engines (OL-Services, based on the Python engine in the current prototype). OL-UI is a web service simulating a desktop and a window manager. The Code is stored on the specialized data spaces - notebooks, which are the documents processed by OL-Services.

Authors present the user's query processing as follows:

- OL-Core and OL-Services are included in the service.
- One notebook represents one computational task.
- The system may recommend notebooks to the other users
- The content of the notebooks is annotated with ontology

The application classifies the user's queries (computational tasks) and reserves communication services of the system in order to guarantee the Quality of Experience (QoE) for the user. The computational tasks are scheduled to reduce the waiting time of the user by monitoring and linking to the IPv6 QoS infrastructure.
General tasks of Online Lab presented in article:

- to compose a computational service, provided that the request stream from the users is known or predicted
- possibility of implementing the dedicated computational services to other applications.

3.3 Social Network

E-mail Network at WUT

The Authors of [83] describe e-mail network at Wrocław University of Technology (WUT). Experiments presented were carried out on the logs from the WUT mail server. The Authors have found that, despite significant changes in the networks structure, the statistical distribution of the subgraphs remains stable, which led to the idea of characterizing network dynamics by the evolutionary patterns of the subgraphs. They assume that prediction of the rapidly changing social network observed over a short period of time should result from the analysis of dependencies and correlations of the activity of the nodes, from the level of the samples of three patterns and the connection between triple of nodes. Introduced is a model that bases on a mixture of Markov chains, which gives very a promising outcome (the mean error rate at the level of approx. 8%). Presented approach groups triad trajectories info clusters. This model was trained on data by expectation-maximization algorithm.

Link Prediction in PlaTel

In [84] link prediction problem in Dynamic Network of Services was discussed. The concept of Network of Services emerges from the patterns of interactions resulting from the composition of Web Services. The article presents the concepts of the Network of Services, which are followed by a research on link prediction methods and their accuracy.

The paper refers to the current models by describing service repositories that are used in the SOA systems. Later on, the paper introduces the concept of *Network of Web Services (NoS)*. NoS is build with all the services in systems service repository and represents all the possible parameter transfers between the services. After introducing the concept of NoS, as a result from service composition and execution, the Authors propose a general approach to characterization of dynamic patterns of interaction between the Web services. This approach uses NoS to represent service activity in a given time window. The Authors introduce the concept of *Dynamic Network of Service s (DNoS)* used to store information about time-dependent parameters interchange between the services. They also notice that this representation is analogous to the representation of dynamic social networks of interactions between humans. Based on this analogy, the Authors propose a web service usage analysis methodology. After suggesting all the concepts for the further research, they present the PlaTel (Platform for ICT solutions planning and monitoring) framework (supporting business processes in distributed ICT environment based on SOA paradigm), used in their research.

The main part of the article describes the experimental comparison of three link prediction algorithms: Preferential Attachment (PA), Common Neighbours (CN) and Triad Transition Matrix (TTM) applied to assess the future service usage and structure of the resulting DNoS. After presenting the experiments assumptions and their methodology, the Authors discussed the results. In conclusion, they stated that the evolution of *DNoS* is not driven by social evolutionary scheme and that the predictors using time series analysis, subgraph structure mining and network statistics, would perform better when making use of dynamic networks of services. The concepts presented in the paper are novel, and thus, they provide the basis for further research.

4 Summary

In this article, we surveyed recent developments and application in Service Science. At the beginning, we explained terms like Service Science, Service Systems as well as two paradigms of Service Oriented Computing and SOA. Also, we presented some challenges of Service Science in section 1.3. Three directions: Service Composition, Knowledge Engineering and Resource Allocation, were discussed. We introduced the present state-of-the art in each of the fields supported by the representative example applications in three groups: eHealth systems, eLearning system and Social Network.

References

1. Barile, S., Polese, F.: Smart Service Systems and Viable Service Systems: Applying Systems Theory to Service Science. Smart Service Systems and Viable Service Systems Service Science 2, 21–40 (2010)
2. Bell, D.: The Coming of Post-Industrial Society. The Educational Forum 40(4), 574–579 (1976)
3. IBM, http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/servicescience/ (accessed May 2013)
4. Maglio, P., Spohrer, J.: Fundamentals of service science. Journal of the Academy of Marketing Science 36, 18–20 (2008)
5. Spohrer, J., Maglio, P., Gruhl, D.: Steps toward a science of service systems. Computer 40, 71–77 (2007)
6. Clavier, P., Lotriet, H., van Loggerenberg, J.: Business Intelligence Challenges in the Context of Goods- and Service-Dominant Logic, Maui, HI, pp. 4138–4147 (2012)
7. Vargo, S., Lusch, R.: Evolving to a New Dominant Logic for Marketing. Journal of Marketing 68, 1–17 (2004)
8. Vargo, S., Lusch, R.: Service-dominant logic: continuing the evolution. Journal of the Academy 36, 1–10 (2008)

9. Vargo, S., Lusch, R.: The Service-Dominant Logic of Marketing: Dialog, Debate, and Directions. M.E. Sharpe, Armonk (2006)
10. Spohrer, J., Kwan, S.: Service Science, Management, Engineering, and Design (SSMED): An Emerging Discipline - Outline & References. Journal of Information Systems in the Service Sector (IJISSS) 1(3), 39 (2009)
11. Maglio, P., Srinivasan, S., Kreulen, J., Spohrer, J.: Service systems, service scientists, SSME, an innovation. Communications of the ACM 49, 81–85 (2006)
12. Ng, I.C.L., Maull, R.: Embedding the New Discipline of Service Science: A Service Science Research Agenda. In: Powell, L., Shi, L., Warren, B. (eds.) IEEE International Conference on Service Operations, Chicago (2009)
13. Papazoglou, M., Georgakopoulos, D.: Service-Oriented Computing. Communication of the ACM 46(10), 25–28 (2003)
14. Qiu, R., Fang, Z., Shen, H., Yu, M.: Towards service science, engineering and practice. International Journal of Services Operations and Informatics 2(2), 103–113 (2007)
15. Spohrer, J., Anderson, L., Pass, N., Ager, T.: Service Science e Service Dominant Logic. Otago Forum 2, 4–18 (2008)
16. Spohrer, J., Vargo, S., Maglio, P., Caswell, N.: The service system is the basic abstraction of service science. In: HICSS Conference (2008)
17. Vargo, S.L., Maglio, P.P., Akaka, M.A.: On value and value co-creation: a service systems and service logic perspectiv. European Management Journal 26(3), 145–152 (2008)
18. Alter, S.: Challenges for Service Science. Journal of Information Technology Theory and Application (JITTA) 13(3), Article 3 (2012)
19. Rao, J., Su, X.: A Survey of Automated Web Service Composition Methods. In: Cardoso, J., Sheth, A.P. (eds.) SWSWPC 2004. LNCS, vol. 3387, pp. 43–54. Springer, Heidelberg (2005)
20. Rao, J., Su, X.: A Survey of Automated Web Service Composition Methods. In: Cardoso, J., Sheth, A.P. (eds.) SWSWPC 2004. LNCS, vol. 3387, pp. 43–54. Springer, Heidelberg (2005)
21. Jong Myoung, K., Chang Ouk, K., Ick-Hyun, K.: Quality-of-service oriented web service omposition algorithm and planning architecture. The Journal of Systems and Software 81(11), 2079–2090 (2008)
22. Stelmach, P., Świątek, P., Falas, Ł., Schauer, P., Kokot, A., Demkiewicz, M.: Planning-Based Method for Communication Protocol Negotiation in a Composition of Data Stream Processing Services. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2013. CCIS, vol. 370, pp. 531–540. Springer, Heidelberg (2013)
23. Fraś, M., Grzech, A., Juszczyszyn, K., Kołaczek, G., Kwiatkowski, J., Prusiewicz, A., Sobecki, J., Świątek, P., Wasilewski, A.: Smart Work Workbench; Integrated Tool for IT Services Planning, Management, Execution and Evaluation. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 557–571. Springer, Heidelberg (2011)
24. Grzech, A., Świątek, P.: Complex Services Availability in Service Oriented Systems. In: 2011 21st International Conference on Systems Engineering (ICSEng), pp. 227–232 (2011)
25. Grzech, A., Rygielski, P., Świątek, P.: Translations of service level agreement in systems based on service-oriented architectures. Cybernetics and Systems: An International Journal 41(8), 610–627

26. Grzech, A., Świątek, P.: Modeling and optimization of complex services in service-based systems. Cybernetics and Systems: An International Journal 40(8), 706–723
27. Rygielski, P., Świątek, P.: Graph-fold: an efficient method for complex service execution plan optimization. Service Science 38(3), 25–32 (2010)
28. Klusch, M., Fries, B., Sycara, K.: OWLS-MX: A hybrid Semantic Web service matchmaker for OWL-S services. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 121–133 (2009)
29. Cena, F., Furnari, R.: Discovering and Exchanging Information about Users in a SOA. Communication of SIWN - Systemics and Informatics World Net 4(3), 34–38 (2008)
30. Karakoc, E., Senkul, P.: Composing semantic Webservices under constraints. Expert Systems with Applications, 11021–11029 (2009)
31. Falas, Ł., Stelmach, P.: Web Service Composition with Uncertain Non-functional Parameters. In: Camarinha-Matos, L.M., Tomic, S., Graç, P. (eds.) Technological Innovation for the Internet of Things, Costa de Caparica, pp. 45–52 (2013)
32. Wiesemann, W., Hochreiter, R., Kuhn, D.: A Stochastic Programming Approach for qosaware Service Composition. In: Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid, Lyon (2008)
33. Hwang, S., Wang, H., Tang, J., Srivastava, J.: A Probabilistic Approach to Modeling and Estimating the QoS of Web-services-based Workflows. Journal of Information Sciences (INS) 177(23), 5484–5503 (2007)
34. Bhowan, U., Johnston, M., Zhang: Developing New Fitness Functions in Genetic Programming. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42(2), 406–421 (2011)
35. Tomczak, J.M., Cieślińska, K., Pleszkun, M.: Development of Service Composition by Applying ICT Service Mapping. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2012. CCIS, vol. 291, pp. 45–54. Springer, Heidelberg (2012)
36. Kohavi, R.: The Power of Decision Tables. In: Lavrač, N., Wrobel, S. (eds.) ECML 1995. LNCS, vol. 912, pp. 174–189. Springer, Heidelberg (1995)
37. Pawlak, Z.: Rough set theory and its applications. J. of Telecom. Inf. Tech. 3, 7–10 (2002)
38. Świątek, P., Stelmach, P., Prusiewicz, A., Juszczyszyn, K.: Service composition in knowledge-based SOA systems
39. Feigenbaum, E., McCorduck, P.: The fifth generation, 1st edn. Addison-Wesley, Reading (1983)
40. Prusiewicz, A., Zięba, M.: On some method for limited services selection. International Journal of Intelligent Information and Database Systems 5(5), 493–509 (2011)
41. Prusiewicz, A., Stelmach, P.: An improved method for services selection. In: Grzech, A., Świątek, P., Brzostowski, K. (eds.) Applications of Systems Science. Exit, Warsaw (2010)
42. Zięba, M., Świątek, J.: Various methods of combining classifiers for ensemble algorithms. Applications of System Science 91(1), 81 (2010)
43. Tomczak, J., Zięba, M.: On-line bayesian context change detection in web service systems. In: HotTopiCS 2013. Proceedings of the 2013 International Workshop on Hot Topics in Cloud, pp. 3–10 (2013)
44. Prusiewicz, A., Zięba, M.: Services Recommendation in Systems Based on Service Oriented Architecture by Applying Modified ROCK Algorithm. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010, Part II. CCIS, vol. 88, pp. 226–238. Springer, Heidelberg (2010)
45. Huang, G., Tian, Y., Chang.: A knowledge representation architecture for remote sensing image understanding systems. In: 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC), vol. 1, pp. 202–205 (2011)

46. Hu-Chen, L., Long, L., Qing-Lian, L., Nan, L.: Knowledge Acquisition and Representation Using Fuzzy Evidential Reasoning and Dynamic Adaptive Fuzzy Petri Nets. IEEE Transactions on Cybernetics 43(3), 1059–1072 (2013)
47. Świątek, P., Grzech, A., Rygielski, P.: Adaptive packet scheduling for requests delay guaranties in packet-switched computer communication network. Systems Science 36(1), 7–12 (2010)
48. Świątek, P., Drwal, M., Grzech, A.: Providing strict QoS guaranties for flows with time-varying capacity requirements. In: 21st International Conference on Systems Engineering (ICSEng), pp. 279–284 (2011)
49. Grzech, A., Świątek, P.: The influence of load prediction methods on the quality of service of connections in the multiprocessor. Systems Science 35(3), 7–14 (2009)
50. Tomczak, J.M.: On-line change detection for resource allocation in service-oriented systems. In: Camarinha-Matos, L.M., Shahamatnia, E., Nunes, G. (eds.) DoCEIS 2012. IFIP AICT, vol. 372, pp. 51–58. Springer, Heidelberg (2012)
51. Rygielski, P., Tomczak, J.M.: Context change detection for resource allocation in service-oriented systems. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS, vol. 6882, pp. 591–600. Springer, Heidelberg (2011)
52. Sebastião, R., Gama, J., Rodrigues, P.P., Bernardes, J.: Monitoring Incremental Histogram Distribution for Change Detection in Data Streams. In: Gaber, M.M., Vatsavai, R.R., Omitaomu, O.A., Gama, J., Chawla, N.V., Ganguly, A.R. (eds.) Sensor-KDD 2008. LNCS, vol. 5840, pp. 25–42. Springer, Heidelberg (2010)
53. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York (2009)
54. Song, Y., Sun, Y., Shi, W.: A Two-Tiered On-Demand Resource Allocation Mechanism for VM-Based Data Centers. IEEE Transactions on Services Computing 6(1), 116–129 (2013)
55. Świątek, P., Rygielski, P., Juszczyszyn, K., Grzech, A.: User Assignment and Movement Prediction in Wireless Networks. Cybernetics and Systems: An International Journal 43(4), 340–353 (2012)
56. Grzech, A., Świątek, P.: Parallel processing of connection streams in nodes of packet-switched computer communication systems. Cybernetics and Systems: An International Journal 39(2), 155–170 (2008)
57. Varshney, U.: Pervasive healthcare and wireless health monitoring. Mobile Networks and Applications 12, 113–127 (2007)
58. Meingast, M., Roosta, T., Sastry, S.: Security and privacy issues with health care information technology. In: Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York, pp. 5053–5058 (2006)
59. Halperin, D., Heydt-Benjamin, T., Fu, K., Kohno, T.: Security and privacy for implantable medical devices. Pervasive Computing 7, 30–39 (2008)
60. Grzech, A., Świątek, P., Rygielski, P.: Dynamic Resources Allocation for Delivery of Personalized Services. In: Cellary, W., Estevez, E. (eds.) Software Services for e-World. IFIP AICT, vol. 341, pp. 17–28. Springer, Heidelberg (2010)
61. Guo, L., Zhang, C., Sun, J., Fang, Y.: PAAS: A Privacy-Preserving Attribute-Based Authentication System for eHealth Networks. In: IEEE 32nd International Conference on Distributed Computing Systems, ICDCS (2012)
62. Lin, X., Lu, R., Shen, X., Nemoto, Y., Kato, N.: Sage: A Strong Privacy-Preserving Scheme Against Global Eavesdropping for ehealth Systems. IEEE Journal on Selected Areas in Communications 27(4) (2009)

63. Świątek, P., Juszczyszyn, K., Brzostowski, K., Drapała, J., Grzech, A.: Supporting Content, Context and User Awareness in Future Internet Applications. In: Álvarez, F., et al. (eds.) FIA 2012. LNCS, vol. 7281, pp. 154–165. Springer, Heidelberg (2012)
64. Brzostowski, K., Drapała, J., Grzech, A., Świątek, P.: Adaptive Decision Support System for Automatic. Cybernetics and Systems: An International Journal 44(23), 204–221 (2013)
65. Tomczak, J., Gonczarek, A.: Decision rules extraction from data stream in the presence of changing context for diabetes treatment. Knowledge and Information Systems 34(3), 521–546 (2013)
66. Świątek, J., Brzostowski, K., Tomczak, J.: Computer aided physician interview for remote control system of diabetes therapy. In: Józefczyk, J., Lasker, G., Tecumseh (eds.) 23rd International Conference on System Research, Informatics and Cybernetics, Baden-Baden, Germany, pp. 8–13 (August 2011)
67. Harries, M.: Splice-2 comparative evaluation: electricity pricing. Technical Report UNSW-CSE-TR-9905
68. Kahn, M.: UCI Machine Learning Repository, `http://archive.Ics.uci.edu/ml/datasets/Diabetes`
69. Kukla, E., Nguyen, N., Sobecki, J., Danilowicz, C., Lenar, M.: A model conception for optimal scenario determination in an intelligent learning system. ITSE -International Journal of Interactive Technology and Smart Education 1(3), 171–184 (2004)
70. Kozierkiewicz-Hetmańska, A.: A method for scenario recommendation in intelligent e-learning systems. Cybernetics and Systems: An International Journal 42(2), 82–99 (2011)
71. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: Proc. 15th Conf. Uncertainty in Artificial Intelligence, pp. 289–296 (1999)
72. Rocchio, J.: SMART Retrieval System Experiments in Automatic Document Processing. In: Salton G. (ed.) Relevance Feedback in Information Retrieval, ch. 14. Prentice-Hall (1971)
73. Shardanand, U., Maes, P.: Social Information Filtering: Algorithms for Automating Word of Mouth. In: Proc. Conf. Human Factors in Computing Systems (1995)
74. Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. Communications of the ACM 40(3), 66–72 (1997)
75. Adomavicius, G., Tuzhilin, A.: Towards the Next Generation of Recommender Systems: A Survey ofthe State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
76. Herlocker, J., Konstan, J., Riedl, J.: Explaining Collaborative Filtering Recommendations. In: Proceedings Conf. Computer Supported Cooperative Work. ACM (2000)
77. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval Journal 4(2), 133–151 (2001)
78. Sarwar, B., Karypis, Konstan, J., Riedl, J.: Item-Based Collaborative Filtering Recommendation Algorithms. In: Proc. 10th Int'l. WWW Conf. (2001)
79. Ramakrishnan, N., Keller, B., Mirza, B., Grama, A., Karypis, G.: Privacy Risks in Recommender Systems. IEEE Internet Computing 5(6), 54–62 (2001)
80. Sobecki, J., Tomczak, J.M.: Student Courses Recommendation Using Ant Colony Optimization. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010, Part II. LNCS (LNAI), vol. 5991, pp. 124–133. Springer, Heidelberg (2010)
81. Tomczak, J.M., Świątek, J.: Personalisation in Service-Oriented Systems Using Markov Chain Model and Bayesian Inference. In: Camarinha-Matos, L.M. (ed.) DoCEIS 2011. IFIP AICT, vol. 349, pp. 91–98. Springer, Heidelberg (2011)

82. Tomczak, J., Świątek, J., Brzostowski, K.: Bayesian classifiers with incremental learning for nonstationary datastreams. In: Grzech, A., Świątek, P., Drapała, J. (eds.) Advances in Systems Science. Exit, Warsaw (2010)
83. Juszczyszyn, K., Gonczarek, A., Tomczak, J., Musiał, K., Budka, M.: A Probabilistic Approach to Structural Change Prediction in Evolving Social Networks. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 996–1001 (2012)
84. Grzech, A., Juszczyszyn, K., Stelmach, P., Falas, Ł.: Link prediction in dynamic networks of services emerging during deployment and execution of web services. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS, vol. 7654, pp. 109–120. Springer, Heidelberg (2012)

Toward Self-adaptive Ecosystems of Services in Dynamic Environments

Francisco Cervantes[1], Michel Occello[1], Félix Ramos[2], and Jean-Paul Jamont[1]

[1] Université de Grenoble, LCIS, F-26900, Valence, Cedex 9, France
{francisco.cervantes,michel.occello,jean-paul.jamont}@lcis.grenoble-inp.fr
[2] CINVESTAV del IPN, Unidad Guadalajara, Zapopan 45010, Jalisco, México
framos@gdl.cinvestav.mx

Abstract. Dynamic composition of services enables users to use complex services with a minimal human intervention. Services that interact in highly dynamic physical world can not remain static, they must be continuously adapted to changes in their environment. Classical mechanisms of service composition are not suitable to implement the service adaptation in open and dynamic environments. We propose an approach based on Multiagent Systems to develop services ecosystems and face the challenge of service adaptation as a constraint satisfaction problem. In this paper, effects of service dispersion and density in the ecosystem are showed.

Keywords: multiagent systems, services, ecosystem, self-adaptation.

1 Introduction

An interesting current topic in smart environments domain is the growing importance of Service Oriented Computing (*SOC*) paradigm in the development of enabled systems to provide pervasive services that interact with the physical world. Additionally, with the evolution of complex services has emerged the need to build autonomous devices with teamwork skills. SOC and Multiagent Systems (*MAS*) are two knowledge domains suitable to complement each other to face the challenge of service adaptation in dynamic environments. On one hand, pervasive systems based on *SOC* represent a domain where MAS can perform significant contributions. Pervasive systems with services that depend on the physical world (open and dynamic), are characterized by a continuous evolution. For example, resources and services embedded in mobile devices can be no more available, new services can be enabled, and functional requirements may change over time. This challenges can be faced by $MAS's$ concepts such as autonomy and cooperation. On the other hand, the service's features such as soft coupled, dynamic selection and composition enable a natural introduction of the *MAS* approach.

Several authors have proposed to use the multi-agent systems paradigm to develop adaptive services [1] [2] [3]. Whereas [1] [2] [3] are efforts in designing

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_64,

and building enriched services (with own goals and teamwork skills) for supporting simulations of software-services adaptations, this present research is an early effort, that introduces a self-adaptive ecosystem of services based on MAS paradigm for dealing with dynamism on the physical environment. In natural ecosystems, members are organized by species and work to achieve common objectives in a shared physical environment [7]. An ecosystem of services consists of two elements: participants (e.g., devices) that can provide services and an environment [4]. Each participant can be viewed as an individual or an organization and has its own role to play. In the ecosystem, participants work together to achieve their objectives and the environment supports the participant's work. The ecosystem adaptation emerges from the local interactions of its participants. In the literature we can found works about services' ecosystems, for example [5] [6]. However most of the works focus on business-oriented services (often these approaches not consider the interaction of services with the physical world) or propose to use some non-suitable techniques for use online such as genetic algorithms.

We propose to use the approach of Multiagent Systems (MAS) to face an open problem: service adaptation in dynamic environment using resources in the user's vicinity. The objective is to automate the adaptation process based on environment changes, specifically in local resources (availability of services, devices, etc.) and current functional requirements. In this paper, a framework for self-adaptive ecosystems of services to be deployed in dynamic environment is presented. Each participant and source of the system is instantiated by an agent. The ecosystem will merge the traditional service composition with the service adaptation. Ecosystem's self-adaptation is supported by two techniques: (i) Clustering based on skills, location and aims is used to build dynamic clusters, and (ii) Dynamic constraints satisfaction for dynamic adaptations of services. The novel features of this work are:

- Designing a framework for self-adaptive ecosystems of services (Section 2).
- Applying (skills, location, aims)-based clustering and constraint satisfaction techniques with MAS to build self-adaptive ecosystems; distributed and mobile clustering (Section 3) to cope with the openness and dynamism of physical environment, and a constraint satisfaction model(Section 4) to dynamically adapt services in the ecosystem.
- Implementing virtual and physical participants' behaviors for self-adaptation into a MAS-based test bed and experiments (Section 5) to evaluate the performance of the MAS-based solution.

In section 6 we present conclusions and future work.

2 A Framework for Self-adaptive Ecosystems of Services

We propose an architecture where participants are organized based on their skills, location and current aims. The MAS-based architecture for self-adaptive ecosystem of services consists of participant agents, services and a service ontology. A participant can perform two self-imposed roles; provider and/or customer, and it can represent devices, users, resources or applications. This architecture is abstract and simple because we think that the organization must emerge of local interactions along time and space.

2.1 Participant Agents

A participant agent (PA) is an autonomous system that resides on a device, and that can sense and change its environment. We represent a PA as a tuple $(Skills, Aims, Behaviors)$.

Skills of a PA are defined by a set of device's features (such as sensors, actuators, storage capacity, communication interfaces) and a set of services that it can provide to others PAs.

Aims represent the potential set of functional requirements that a PA will need to fulfill (for the shake of simplicity, in this paper we assume a finite and fixed set of aims). We have defined an ontology of potential aims and skills that will be used in the clustering process.

Behaviors define how a PA perceives and modifies its environment. A behavior definition has two elements: preconditions and actions.

Any PA carrying out an action in support of others PAs is providing a service. A service can be atomic (atomic service is a service residing on a single device and has owns all required decision capabilities) or composed. In this work, we distinguish two types of atomic services that an device can provide in the physical world.

- Software-service, which represents a logical solution to process and manage information.
- Hardware-service, which represents an action over the physical world trough hardware such as sensors and actuators.

Using atomic services we can provide composed services constituted by software services and hardware-services. Each service description is based on a shared ontology service. Service functionality is represented by the requirements that it fulfills, such as, sensing temperature. Inputs of the service are represented by the requirement's parameters, e.g., user location. The service's output is an action that contribute to fulfill a functional requirement.

2.2 *PA*'s Roles

In the required process, to perform a service's composition and adaptation, a PA can choose to play three roles; provider, customer or representative. Acquired roles are the result of PA's behaviors, environment state and interactions with others PAs.

Provider Role. This role is acquired when a PA has accepted an obligation to carry out an action in support of an other PA (PA has acquired an obligation to provide a service). We use PPA to indicate that the PA is performing a provider role. The behavior of this role depends of the kind of requested service.

- If the requested service is atomic then the PPA executes the actions required to fulfill the service and only interact with others PAs to request resources (if more resources are necessary).
- If the requested service is composed then the PPA must interact with others PAs to request services needed to fulfill the composed service. In this regard, a PPA can be seen as the owner of a temporal virtual ecosystem (devices with a common goal in a shared environment along a finite time period), where some PA are registered as members. The main function of a PPA is to delegate functional requirements to others PAs. A PPA provides common and collaborative contexts to PAs, i.e. PAs managed by the same PPA are teammates that share a common goal. This allows PAs to delegate functional requirements among each other, promoting the service composition and adaptation.

Customer role represents a PA that needs of other PAs to fulfills a functional requirement, we represent this role as CPA and its behaviors are: (i) formalizing functional requirements by making use of the service ontology, (ii) decomposing composed requirements into atomic requirements, (iii) contacting one ore more PA to carry out actions to fulfills the requirements, (iv) combining PA's results into a single composed service. PAs self-acquire obligations (to provide services) to promote the composition and adaptation of their own needed services.

Representative role represents a PA with the best skills that characterize a PA's cluster. We use RPA to indicate that the PA is performing a representative role. RPA is used to build initial clusters of PAs. It is based on the PA's skills and allows PAs to find other PAs with similar skills (we can perform an analogy of these initial clusters and the concept of organization based on species in a natural ecosystem). Initial clustering is based on the hypothesis: PAs with similar skills have similar aims.

In the (Fig. 1) we show the conceptual framework to build services' ecosystems based on a MAS approach with all the components defined before. This architecture supports the ecosystem's life-cycle required to provide services; organization (based on skills and aims), composition and adaptation.

3 *PA*s' Dynamic Clustering

Based on the idea; members of same species work together to achieve common goals. We propose to build clusters of PAs (for the shake of simplicity, each PA represents a device) to provide services and achieve common aims. However, physical environment and functional requirements are constantly changing,

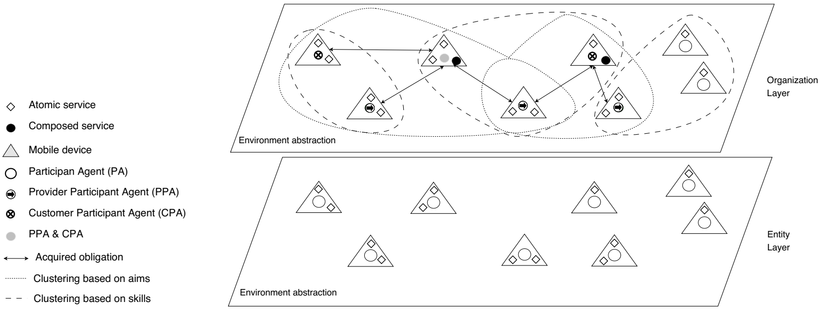


Fig. 1. Architecture from a conceptual point of view of a Ecosystem of Services

therefore it is false that the PAs' clusters in a ecosystem must remain static. For these reason we propose a clustering algorithm based on skills and locations to build initial clusters, where each cluster is adapted guided by current aims of each PA.

3.1 Preliminaries

Every participant is instantiated and represented by an agent, this agent has an unique identifier denoted by PA. Each PA has a set of skills to provide one or more services. In this paper we assume the emergence of clusters based on a minimal percent of similarity in three dimensions; skills, locations and current aims. In the ecosystem of services, this clusters means PA's working to provide services and achieve common aims. However, this clusters are not fixed and may change along the time.

In this paper, it is out of our scope to derive similarity functions. In regard of this, for the shake of simplicity, we assume the existence of three functions; (i) a similarity function based on PA's skills that enable PAs to determine distances among them, (ii) a function to determine a similarity percent between to aims and (iii) a function to determine the geographical distance between two PAs. Each cluster in the ecosystem should have a representative participant agent (RPA) and some PAs. The choice of the RPA is based on the skills associated to each PA and location; the PA with best skills to represent a cluster play the RPA role. In order to achieve an appropriate partition of initial clusters in the ecosystem, the clustering process must satisfy the following properties:

- Every PA has at least a RPA as a neighbor (two PAs are neighbors if they belong to the same cluster).
- Every PA must affiliate with the neighboring RPA that has greater similarity to it, based on skills, location and current aims.
- Two $RPAs$ cannot be neighbors.

The clustering process is executed in all PAs and each one decides its own role (PPA, RPA or PA); depending only on the neighbors' decisions. Thus initially, only the PA with best skills will broadcast a message to its neighbors stating that he will be the RPA. When one or more of these messages are received, the PA will choice to join the cluster of the RPA with the biggest skills. If any message has been received by PA from a PA with higher skills, then PA will send a message to promote it-self as a new RPA.

3.2 Clustering Process

The creation and adaptation process of clusters, in the ecosystem, is driven by messages; a specified behavior will be executed at the *PAs* depending on the reception of the corresponding message. The main messages used by PA in the clustering process are:

- "*ChangeRPA*" is used by a PA to inform its neighbors that it is going to be the RPA.
- "*GoInto*", with which a PA communicates to its neighbors that it will be part of a cluster whose RPA is a neighbor.
- *CurrentAim* is used by a PA to inform its neighbors that it has a new aim. If neighbors' aims are similar to his current aim, then PA remains in the current cluster. Otherwise PA find, in his vicinity, another RPA with similar goals to his current aim. A PA may have one or more current aims, it can thus also belong to more than one cluster.
- *AddRPA* is used by a PA to inform its neighbors that it will be part of other clusters, but remain in the current cluster too.
- *NewPA* is used by a PA that has come to the ecosystem. This messages start the clustering process based on three dimensions: skills, location and aims. The priority of each dimension can be adjusted depending of the problem domain.
- *LostPA* is sent when a PA detects a link failure with another PA (we assume the existence of a function of low level to detect link failures). This messages promote a local adaptation of the cluster.

All clusters in the ecosystem are dynamic and we distinguish two types: primary and secondary. A primary cluster is based on similarity skills, while an secondary cluster is based on current common aims. Each PA can belong to an cluster's primary and several secondary clusters.

4 Service Composition and Adaptation

In order to achieve the dynamic composition and adaptation of services, we model the process as a constraint satisfaction problem and each PA adopt the well-known asynchronous backtracking (ABT) algorithm [8]. Using constraints allows us to perform the composition process to achieve the service adaptation with the same algorithm based on constraints updates.

4.1 Service Request

Providing a service at the user layer, usually requires the composition or adaptation of services (at organizational layer) offered in the user's vicinity. In this section we use the term *task* to refer to a service at the user layer. From a $task_x$ request is generated a functional requirement description (FRD_x). FRD_x is divided into required services (RS_x) and service constraints (SC_x). These constraints can be logical and/or physical. A RS_x represents a set of services (from organizational and entity layers) required to fulfill the requested $task_x$.

4.2 *CSP* Model

In the service ecosystem can be several requested $tasks$, therefore we need to fulfill several $FRDs$. We can model the service composition and adaptation as a constraint satisfaction problem (CSP). We are using the distributed model presented by [8] to fulfill each FRD_x in the ecosystem. Each FRD_x is a CSP represented by a tuple (RS_x, SC_x ,SD_x). RS_x is a set of needed services $\{rs_1, rs_2, ..., rs_n\}$ to performs the requested $task_x$ (in other words, is the set of problem's variables). SC_x is the set of constraints on RS_x, and SD_x is a set of service domains $\{sd_{rs_1}, sd_{rs_2}, ..., sd_{rs_m}\}$. Each service domain sd is a set of atomic services that will be used to instantiate a variable rs_i. Therefore, service composition and adaptation consists in finding a solution for FRD_x.

The composition and adaptation problem of FRD_x can be represented with a bipartite graph G (Fig. 2) comprised of N required services rs_i and M participant agents (PA_j). A edge between two nodes in G represents a relation that means RS_i can be fulfilled by a service provided by PA_j.

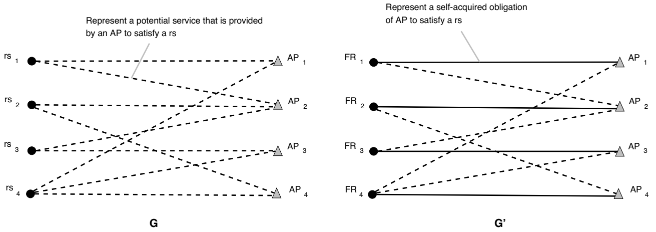

Represent a potential service that is provided by an AP to satisfy a rs
rs 1
rs 2
rs 3
rs 4
AP 1
AP 2
AP 3
AP 4
G
Represent a self-acquired obligation of AP to satisfy a rs
FR 1
FR 2
FR 3
FR 4
AP 1
AP 2
AP 3
AP 4
G'

Fig. 2. Problem represented as a bipartite graph

4.3 Candidates *PA*'s Formation

The PAs' formation is the first step to create the graph G. When a FRD_x has been created the corresponding CPA sends the service request to its neighbors. Using the FRD_x, each PA decide whether or not it can contribute to satisfy some rs_i. Each PA that can contribute to satisfy an rs_i send a PA – *Reply* message to the CPA initiator. The CPA initiator uses thee replies to compute a candidate table containing the addresses of all responding nodes. The nodes in the candidate table are organized in descending priority order based on their skills, location and aims. One the table is formed, the CPA initiator distributes it to all the PAs (candidates) of the table.

4.4 *ABT* Algorithm

With the complete graph G, we apply the ABT algorithm. The execution of ABT is asynchronous and distributed among all the PA candidates; it begins with all candidates self-acquiring all the services in their respective variable and exchanging *Agree*? messages to find conflicts.

Agree? messages are used to spread information of value assignments among PA candidates; each PA uses this information to build its PA view. A PA view reflects a current partial solution. A partial solution is a subset of the final solution, which is expanded by adding committed variables one by one, until a final solution is reached. PAs candidates may have different partial solutions, although as the algorithm progresses, all PAs candidates work toward a single common PA view (the final solution). When an *Agree*? message arrives to a PA candidate, it compares the partial solution of the message with its PA view. PA searches for any inconsistency. Inconsistencies occur when there are several PAs that want to contribute to the same rs_i, or when some constraint has been violated. If inconsistencies are present, PA_i attempts to resolve this by altering its rs_i contributions. If PA_i is able to achieve consistency, it distributes its new PA view in an Agree? message to all candidates, otherwise a backtrack is necessary.

A backtrack action involves to send inconsistencies messages to PAs participants. An inconsistency message specifies exactly which value in rs_i is considered as not good. Every candidate keeps an inconsistency list, which adds to its constraints set SC. On receiving an inconsistency message, the PA attempts to adapt its respective rs_i values, enabling the sender to reenter a consistent state. If the PA is unable to resolve the inconsistency, then it triggers a backtrack. A no solution occurs when any PA can resolve the inconsistency.

5 Simulations and Analysis

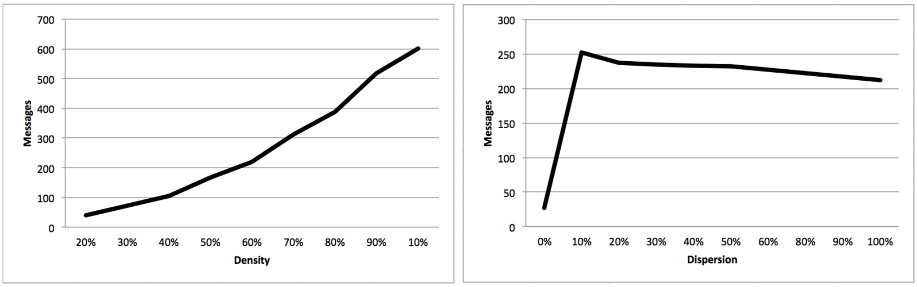
A set of experiments was conducted using the MAS approach defined in Sections 2-4. The test bed was implemented using J-SIM [9], a well-known simulation tool. We use an area of 100 x 100 m containing 25 PAs (each PA represent a mobile device with wireless interface). We set a transmission range of 100 m. All broadcasts follow a bounded hop count of one. The objective of this experiments is to know effects of service density and dispersion on the ecosystem (using the proposed clustering and ABT algorithms).

5.1 Density Effects

We use the concept of service density as the percentage of nodes that have one service required to fulfill an FRD_x. To examine the effects of density on the ecosystem we identify the number of messages in each simulation. The simulation was released for composite lengths of 4 and 7, (composite length is the number of services required to fulfills the FRD_x) and for densities from 20% to 100%. In (Fig. 3a) we show the effect of service density on the number of messages needed to fulfill a FRD_x (values are the average of 100 simulations). Results show that our algorithm suffers from density issues. The effect of a great service density is a robust system but at the same time increases the number of negotiations between PAs. It can affect the battery's life-time of the devices and therefore their autonomy.

5.2 Dispersion Effects

We define service dispersion as the number of PAs required to fulfill a FRD_x. A dispersion at 0% means that all required services to fulfill a FRD_x is in one PA. To examine the effects of dispersion on the ecosystem, we have used the above configuration, but we now set the service density at 50%, and composite length at ten. In (Fig. 3b) we show the effect of services dispersion on the number of messages needed to fulfill a FRD_x (values are the average of 100 simulations). Results show an quick increment in the number of messages when there is services dispersion (without dispersion the number of messages between PAs is minimal) and after the changes in the number of messages required is stable. The results allows us to formulate an early hypothesis: using our algorithm the effect of dispersion on the number of messages is acceptable because with a high percentage of service dispersion we get a system more robust to device's failures.



(a) Density effect

(b) Dispersion effect

Fig. 3. (a) Number of messages used, with respect to the service density for composition lengths 4,7,10. (b)Number of used messages, with respect to the service dispersion for composition lengths 10 and density at 50%

6 Conclusions

We have presented an approach to build an adaptable ecosystem of services based on MAS, service ecosystem, adaptation, clustering (guided by skills, locations and aims) and an asynchronous backtracking algorithm for solving the composition and adaptation problem modeled as a CSP. The approach provides dynamic clusters based on the AP's skills and location (as group of species) and continuously updated by AP's aims. We have also presented a CSP model for the composition and adaptation of services without a central control. Additionally, we have showed and discussed the effects of services density and dispersion on the number of messages between the participant agents. Future work will investigate the design of techniques to achieve a best respond to the service density. However to achieve the deployment of this kind of systems in the physical world we will have to investigate techniques of recursive composition. Additionally, we need to explore the design of techniques to support dynamic constraints (along the composition and adaptation process).

References

1. Gutierrez, O., Sim, K.: Self-Organizing Agents for Service Composition in Cloud Computing. In: 2nd IEEE International Conference on Cloud Computing Technology and Science, pp. 59–66 (2010)
2. Weyns, D., Georgeff, M.: Self-Adaptation Using Multi-Agent Systems. IEEE Software, 86–91 (2010)
3. Torres-Ribero, L.G., Garzon, J.P., Arias-Baez, M.P., Carrillo-Ramos, A., Gonzalez, E.: Agents for Enriched Services (AES): A generic agent - Based adaptation framework. IEEE Colaboration Technologies and Systems, 492–499 (2011)
4. Dong, H., Khadeer, F., Chang, E.: Exploring the Conceptual Model of Digital Ecosystem. In: IEEE Second International Conference on Digital Telecommunications (2007)
5. Briscoe, G., Wilde, P.: Digital Ecosystems: Optimization by a Distributed Intelligence. ArXiv e-prints, Provided by the SAO/NASA Astrophysics Data System (2009)
6. Marín, C.A., Stalker, I., Mehandjiev, N.: Engineering Business Ecosystems Using Environment-Mediated Interactions. In: Weyns, D., Brueckner, S.A., Demazeau, Y. (eds.) EEMMAS 2007. LNCS (LNAI), vol. 5049, pp. 240–258. Springer, Heidelberg (2008)
7. Di-Marzo, G., Gleizes, M.-P., Karageorgos, A.: Self-organising Software - From Natural to Artificial Adaptation. Natural Computing Series. Springer (2011)
8. Yokoo, M., Durfee, E.H., Ishida, T., Kuwabara, K.: Distributed Constraint Satisfaction for Formalizing Distributed Problem Solving. In: Proceedings of the Twelfth IEEE International Conference on Distributed Computing Systems, pp. 614–621 (1992)
9. Hung-ying, T.: Design, realization and evaluation of a component-based compositional software architecture for network simulation. Dissertation at The Ohio State University (2002)

Communication Protocol Negotiation in a Composite Data Stream Processing Service

Paweł Stelmach, Paweł Świątek, and Patryk Schauer

Institute of Computer Science,
Wrocław University of Technology,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
{pawel.stelmach,pawel.swiatek,patryk.schauer}@pwr.wroc.pl

Abstract. Data streaming services gain popularity but still only few works are focused on their composition and its detailed management, like communication protocol negotiation. In this paper, we describe a continued work on a platform for automated composition of distributed data stream processing services. With the platform a process of composite service building is simplified and negotiation among services automated. In the following sections we will present an overview of the platform, describe implemented negotiation approaches and their phases, and finally compare them in an experimental study.

Keywords: Service Oriented Architecture, Data Stream Processing Services, Service Management.

1 Introduction

In todays world much more focus is put on software available in the Internet, especially one that offers its functionality through web services. Such approach gained popularity with the introduction of Service Oriented Architecture (SOA) and SOAP-based web services, that are described by the WS-* stack of standards. Appropriate results on the topic can be found in [18], in which an automated, knowledge-based approach to SOA and services management with a focus on composite services is proposed.

However, with the introduction of video feeds and sensor data, a requirement for a new kind of service has arisen. Such services run constantly in the background processing an ingoing stream of data. Work on the distribution of streaming services can be found in [9], which describes methods for processing sensor data or in [1, 12] introducing specialized middleware for data stream processing.

The natural distribution of source of the stream and its destination was extended via introduction of more computing services in between [1, 8, 11], introducing composite stream processing services in eHealth, rehabilitation and recreation fields. Work in [13] focuses on sportsman and patient monitoring but more examples can be found in computational science and meteorological applications, where there are multiple data sources and multiple recipients interested in the processed data stream [10].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_65, © Springer International Publishing Switzerland 2014

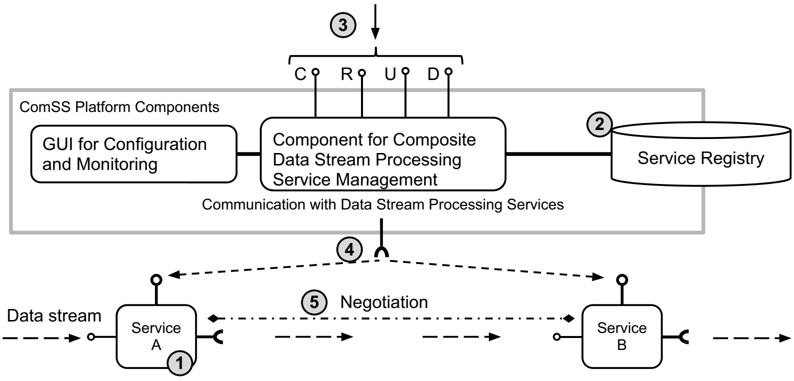

3
C
R
U
D
ComSS Platform Components
GUI for Configuration and Monitoring
Component for Composite Data Stream Processing Service Management
2
Service Registry
Communication with Data Stream Processing Services
4
5
Negotiation
Data stream
Service A
1
Service B

Fig. 1. ComSS Platform overview

In this paper a communication protocol negotiation mechanism is described in relation to the ComSS Platform (COMposition of Streaming Services), which is a result of ongoing work in the area of Future Internet [5, 17] and specialized approaches like ICT services mapping [15], merging or splitting [16] and detecting changes in the QoS [14]. This work introduces and compares multiple approaches to communication protocol negotiation, namely centralized communication planning and two distributed approaches to protocol negotiation: ad-hoc and sequential negotiation.

The ComSS platform offers management over compositions of any data stream processing services. The communication negotiation is of extreme importance and is reflected in negotiation process that takes place during creation of a distributed data stream processing service. Many papers, which take on the subject of composition of services, often refer to services in the WS-* standard [18] or consider stream processing services [2,3], but still perceive them similarly to WS-* type service, omitting their unique characteristics (namely the flexibility in various formats and communication protocols employment).

2 Platform Overview

The main goal of the platform is to manage data stream processing composite services (also called composite streaming services). The designer of a streaming service (also called atomic service) can delegate to the platform all tasks of assembling, disassembling or monitoring of streaming services via the platform API.

The basic scenario for the ComSS Platform is to create a new composite streaming service given a graph of atomic services. It is assumed that those services have been implemented using the provided framework (Fig. 1.1) and are registered in the service registry (Fig. 1.2). The platform searches for appropriate atomic streaming services to fulfil the user composite service request (Fig. 1.3) and forwards them information on neighbour services (Fig. 1.4), with which they will have to negotiate the communication protocol. Streaming services start negotiating (Fig. 1.5) and create new service instances to handle the new composite service request.

Part of the effort to make the atomic streaming service composable, that is managable by the platform, lies with the service designer. He has to follow conventions for the service design and implement necessary libraries for control, negotiation and communication. He can also use the service framework provided with the ComSS Platform and focus on the stream processing algorithm, using the proposed communication and negotiation capabilities of the service framework. In recent works much effort has been put into automation of the negotiation and registration process. The goal for the designer is to, after using the provided framework, register the service in the ComSS Platform repository and its capabilities will be monitored by the platform. Those capabilities refer to the description of service functionality and ever-changing non-functional parameters – like availability, delay etc.. This information is crucial for the ComSS Platform to automatically select services that fulfil non-functional requirements at the moment of the streaming service composition.

3 Negotiation in the Composite Service Creation Process

Data stream processing services are not limited to a single protocol or format – the popular streaming services: video stream processing and sensor data stream processing, have completely different protocols.

The composite service creation process consists of services selection and then their initialisation. Below we focus on the latter, when each service needs to negotiate its communication protocols with other services in the composite service (if the composite service plan determines that they should communicate) and initialise, creating instances for the requested composite streaming service.

The composite service initialisation consists of several stages:

- preparation for communication,
- resource reservation,
- atomic services configuration (preparing instances),
- atomic services (instances) initialisation.

The negotiation usually takes place in service configuration stage or, if the communication is centrally planned, during preparation for communication, the typical negotiation is omitted.

Preparation for communication takes place solely in the ComSS Platform and the remaining stages require some communication among services and the platform.

3.1 Preparation for Communication

The first stage of composite service initialisation is preparation for communication. It decomposes the composite service into elementary service configuration requests, collecting information necessary to communicate with each of the atomic services and initial requirements (like composite service input or output stream format). Based on this decomposition the ComSS Platform can communicate with each service, transferring its initial configuration: its input or output format requirements and neighbouring services, with which the atomic service will communicate. Some service interfaces will be subject to negotiation but the format of some will be forced, whether they are external interfaces of the composite service and connect to the source of destination of the processed stream or simply the platform selected the format centrally instead of the atomic service. The scope of this stage is determined by the negotiation approach - the configuration could be minimal, leaving the negotiation to the atomic services or complete (in planning-based approach), creating a centralized, optimal communication plan and sending its parts to each atomic service.

3.2 Resource Reservation

The goal of this stage is to check the current availability of atomic services resources and reserving resources for new instances of those atomic services. Each atomic service receives following information:

- **composite service identifier.** It allows the atomic service to identify the composite service it is currently a part of. Useful during communication with external world, especially error propagation.
- **negotiation mode.** It defines the behavior of the atomic service during negotiation.
- **list of inputs and outputs.** Not all service interfaces have to be used during processing. The information about inputs and outputs used will determine the following communication negotiation.
- **service initialisation parameters.** Transmitting initialisation parameters to the atomic service allows for defining its behaviour for this particular composite service. An atomic service can implement many algorithms or those algorithms can be parametrized. In such cases it is reasonable to deliver such parameters at initialisation, instead of delivering them in the data stream.

3.3 Atomic Service Configuration

Each atomic service receives following information from the platform:

- **data formats for external interfaces.** This information is sent to atomic services that communicate with external interfaces: sources providing the data stream or destinations consuming it. Those formats are not subject to negotiation and have to be forced upon appropriate atomic services, determining their negotiation.
- **negotiation interface addresses of neighbour atomic services.** For each output the atomic service must receive the address and identification of the negotiation interface of the following atomic service, to which it will transfer the data stream. The atomic service can have multiple outputs,

which data streams are sent to different atomic services, thus the atomic service has to know what is the subject of negotiation between particular neighbouring services.

With the configuration, atomic services begin negotiations with their neighbours. This behaviour is described in more detail in the next section. When each service finalizes its negotiations, it sends an appropriate message to the platform. If all negotiations ended in success, then a composite service can be initialised (all atomic services are initialised).

3.4 Atomic Service Initialisation

The last stage of composite service creation is atomic services initialisation. The ComSS Platform sends to each service a request for its initialisation, which corresponds to creation of an instance with reserved resources and negotiated interfaces.

4 Streaming Service Communication Negotiation Approaches

Ad-hoc Approach. Ad-hoc negotiations are the basic type of negotiations that focus on minimization of communication among services and rely on their initial configuration and first-come-first-served interactions between neighbours. Fig. 2 shows the sequence diagram for this kind of negotiation. Each atomic service starts negotiating with a neighbour service at the same time, according to the approach depicted in the diagram.

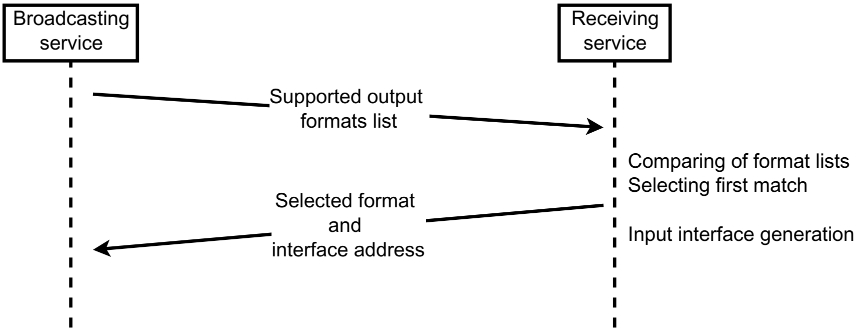

Broadcasting service
Receiving service
Supported output formats list
Comparing of format lists
Selecting first match
Selected format and interface address
Input interface generation

Fig. 2. Sequence diagram for ad-hoc negotiation

This negotiation approach does not guarantee the success of negotiation even if a valid solution exists but can be useful when all atomic services have data stream format convertion capabilities. Preferred formats are selected according to the list of preferred formats supported by a given interface.

Sequential Approach. This method is an extension of the ad-hoc method. This time there is an order to negotiations: from the first atomic service consuming the source data stream to the last atomic service in the composition. The approach is shown in fig. 3 and it allows for renegotiation of formats.

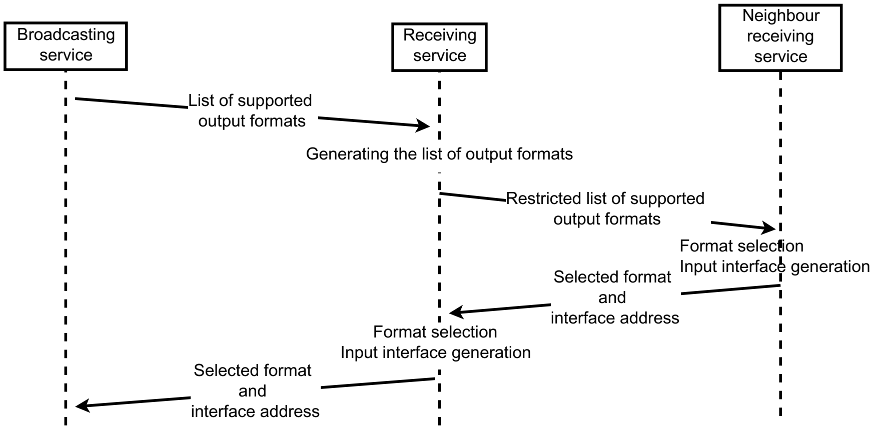

Broadcasting service
Receiving service
Neighbour receiving service
List of supported output formats
Generating the list of output formats
Restricted list of supported output formats
Format selection
Input interface generation
Selected format and interface address
Format selection
Input interface generation
Selected format and interface address

Fig. 3. Sequence diagram for sequential negotiation

The atomic service in a sequence receives a list of formats that a neighbouring service can provide. It filters it with a list of input formats it can receive and generates a list of possible formats it can produce on the output, given the received input formats. If the output formats list in not empty then it is transferred to the next atomic service in the composition. The negotiation continues until one of atomic services in the composition has a format conversion capability or is a last service in a composition. Then it accepts the first valid format from its input list (the last service in a composition has to be able to generate the required destination format) and returns its selection to requesting atomic service. Then, if it is not the last atomic service in a composition, it continues the negotiation with its output formats list.

This procedure guarantees the success of negotiation, provided that a valid solution exists.

Planning-Based Approach. Planning-based approach relies on centralized negotiations plan preparation. The planning takes place during communication preparation, in the ComSS Platform. The algorithm builds a graph of possible format transformations, their costs and communication cost of transferring data in a particular format. The planning-based approach guarantees that an optimal solution will be found if such exists. It is the only approach of the three proposed that faces the communication cost (also other non-functional service and communication parameters). Fig. 4 presents the sequence diagram for the planned negotiations.

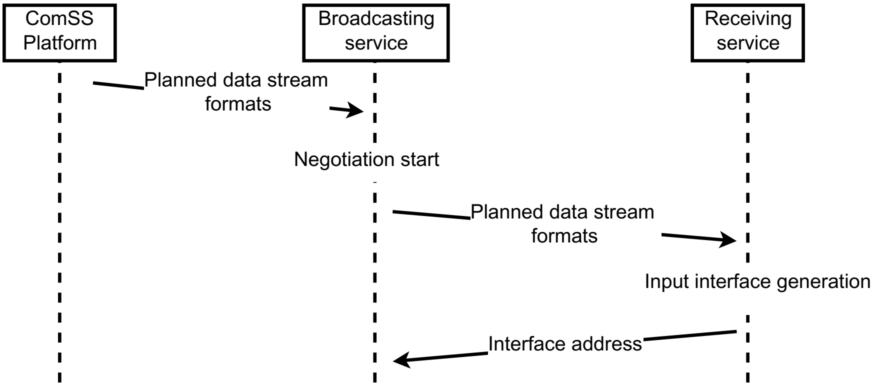


Fig. 4. Sequence diagram for the planning-based approach

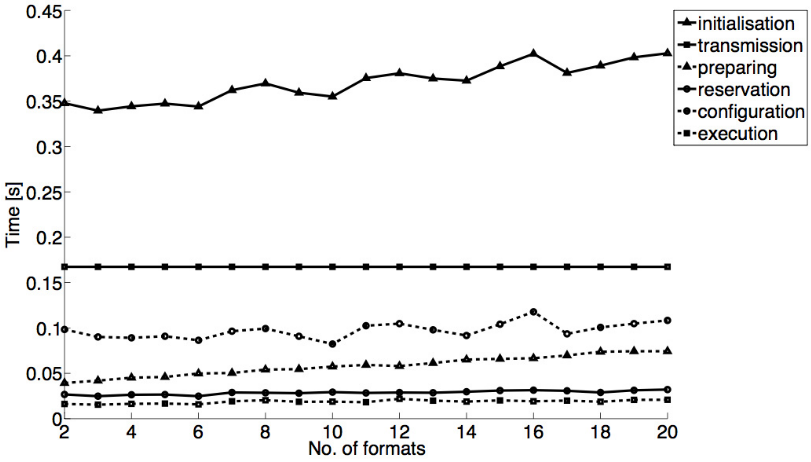


Fig. 5. Composite service initialisation effectiveness for ad-hoc negotiations

5 Experimental Results

5.1 Testing Environment

For testing the composite streaming service initialisation effectiveness of the ComSS Platform a set of different 19 testing graphs was prepared together with appropriate test services. The structure of the graphs was fixed as the graph represented a sportsman monitoring composite service from a practical application scenario of the ComSS Platform. Graphs differed in the numbers of formats in each of the atomic services. For a given graph a number of formats for each interface of each atomic service was fixed. The number of supported formats varied from 2 to 20.

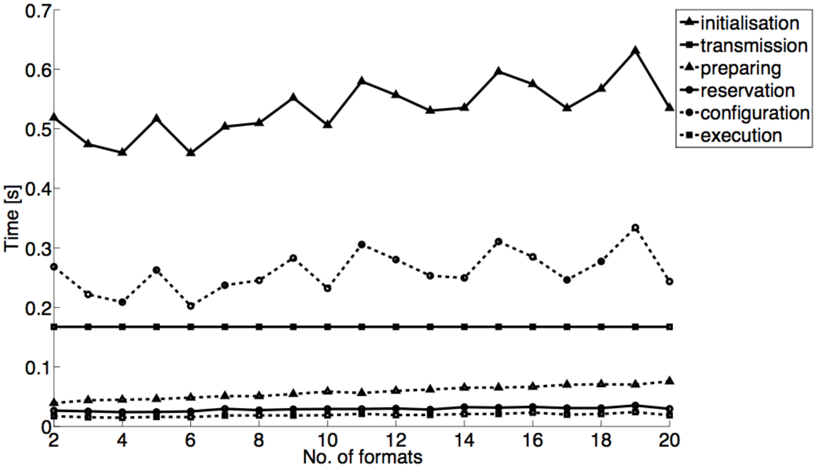


Fig. 6. Composite service initialisation effectiveness for sequential negotiations

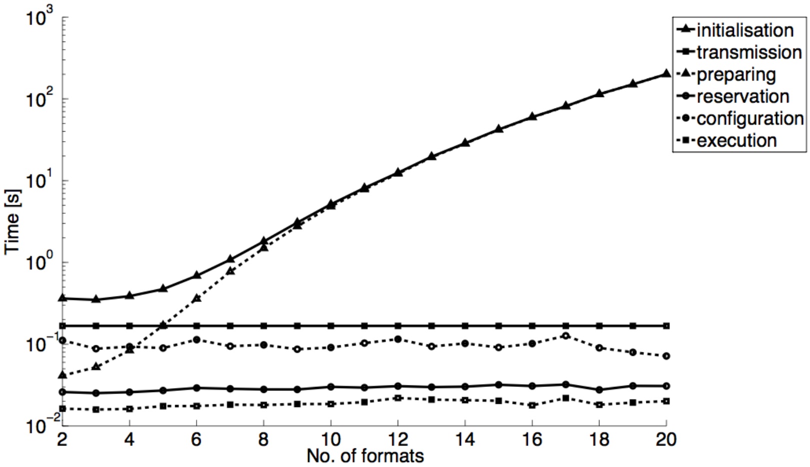


Fig. 7. Composite service initialisation effectiveness for planning-based approach

5.2 Results

Experimental results show that the main costs of initialisation for ad-hoc and sequential approaches are configuration and transmission cost. In sequential approach the configuration cost increases but it is still comparable to the transmission cost. Its greater effectiveness suggests that it should replace the ad-hoc approach. The planning-based approach shows a fast decrease in efficiency with the growing number of supported formats. This result was expected because more formats and convertion capabilities provide more chances to produce a valid

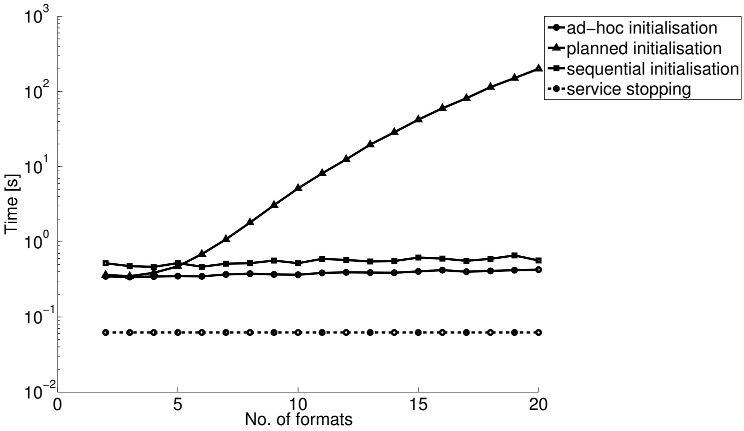


Fig. 8. Comparison of initialisation effectiveness with data transmission time

composite service. However, experiments show that with 5-6 formats per service its efficiency is comparable to the sequential approach and with less formats not only planning provides the optimal solution but its minimal communication allows for greater efficiency.

6 Conclusions and Further Work

Research presented in this paper describes the communication negotiation phase during construction of a composite data stream processing service. A ComSS Platform has been briefly described as an example of a tool that delivers such capability, relaying negotiation requests in accordance with the composite service structure.

Future work will focus on presenting the extended view on ComSS Platform functionality, namely the full process of streaming service composition from user functional and non-functional requirements to executable composite streaming service.

Acknowledgments. The research presented in this paper has been co-financed by the European Union as part of the European Social Fund.

References

1. Chen, L., Reddy, K., Agrawal, G.: Gates: a grid-based middleware for processing distributed data streams. In: High Performance Distributed Computing, pp. 192–201 (June 2004)
2. Riabov, A., Liu, Z.: Planning for Stream Processing Systems. In: Proceedings of the National Conference on Artificial Intelligence (2005)

3. Riabov, A., Liu, Z.: Scalable Planning for Distributed Stream Processing Systems. In: Proceedings of ICAPS (2006)
4. Frossard, P., Verscheure, O., Venkatramani, C.: Signal processing challenges in distributed stream processing systems. In: Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, vol. 5, p. V (May 2006)
5. Grzech, A., Juszczyszyn, K., Świątek, P., Mazurek, C., Sochan, A.: Applications of the future internet engineering project. In: 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel Distributed Computing (SNPD), pp. 635–642 (August 2012)
6. Grzech, A., Rygielski, P., Świątek, P.: Translations of service level agreement in systems based on service-oriented architectures. Cybernetics and Systems 41(8), 610–627 (2010)
7. Grzech, A., Świątek, P., Rygielski, P.: Dynamic resources allocation for delivery of personalized services. In: Cellary, W., Estevez, E. (eds.) Software Services for e-World. IFIP AICT, vol. 341, pp. 17–28. Springer, Heidelberg (2010)
8. Gu, X., Nahrstedt, K.: On composing stream applications in peer-to-peer environments. IEEE Trans. Parallel Distrib. Syst. 17(8), 824–837 (2006)
9. Gu, X., Yu, P., Nahrstedt, K.: Optimal component composition for scalable stream processing. In: Proceedings of the 25th IEEE International Conference on Distributed Computing Systems, ICDCS 2005, pp. 773–782 (June 2005)
10. Liu, Y., Vijayakumar, N., Plale, B.: Stream processing in data-driven computational science. In: 7th IEEE/ACM International Conference on Grid Computing, pp. 160–167 (September 2006)
11. Rueda, C., Gertz, M., Ludascher, B., Hamann, B.: An extensible infrastructure for processing distributed geospatial data streams. In: 18th International Conference on Scientific and Statistical Database Management, pp. 285–290 (2006)
12. Schmidt, S., Legler, T., Schaller, D., Lehner, W.: Real-time scheduling for data stream management systems. In: Proceedings of the 17th Euromicro Conference on Real-Time Systems (ECRTS 2005), pp. 167–176 (July 2005)
13. Świątek, P., Klukowski, P., Brzostowski, K., Drapała, J.: Application of wearable smart system to support physical activity. In: Advances in Knowledge-based and Intelligent Information and Engineering Systems, pp. 1418–1427. IOS Press (2012)
14. Tomczak, J.M., Zięba, M.: On-line bayesian context change detection in web service systems. In: Proceedings of the 2013 International Workshop on Hot Topics in Cloud Services, pp. 3–10 (2013)
15. Tomczak, J.M., Cieślińska, K., Pleszkun, M.: Development of Service Composition by Applying ICT Service Mapping. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2012. CCIS, vol. 291, pp. 45–54. Springer, Heidelberg (2012)
16. Grzech, A., Prusiewicz, A., Zięba, M.: Services merging, splitting and execution in systems based on service oriented architecture paradigm. In: Hippe, Z.S., Kulikowski, J.L., Mroczek, T. (eds.) Human – Computer Systems Interaction, Part 1. AISC, vol. 98, pp. 103–118. Springer, Heidelberg (2012)
17. Świątek, P., Rygielski, P.: Universal comunication platform for QoS-aware delivery of complex services. In: Proceedings of the VIth International Scientific and Technical Conference, pp. 136–139. Publishing House Vezha&Co. (2011)
18. Świątek, P., Stelmach, P., Prusiewicz, A., Juszczyszyn, K.: Service composition in knowledge-based soa systems. New Generation Computing, 165–188 (2012)

Towards a Service-Oriented Platform for Exoplanets Discovery

Paweł Stelmach[1], Rafał Pawłaszek[2], Łukasz Falas[1], and Krzysztof Juszczyszyn[1]

[1] Institute of Computer Science, Wrocław University of Technology,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
{pawel.stelmach,lukasz.falas,krzysztof.juszczyszyn}@pwr.wroc.pl
[2] Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences,
ul. Bartycka 18, 00-716 Warszawa, Poland
pawlaszek@ncac.torun.pl

Abstract. The work in this paper presents the architecture of the service-oriented platform for exoplanets (also known as extrasolar planets) discovery. It is the result of an interdisciplinary cooperation of the three partners with experience in the fields of discovery of extrasolar planets, systems built according to the Service Oriented Architecture paradigm and parallel data processing on graphics cards. The platform architecture presents the integration of distributed, autonomous software components specializing in making observations of space (robotic telescopes), transport of large data over computer networks, distributed data processing on graphics cards, and data visualization. Components offer their functionality through web service interfaces that are invoked in the process coordinated by the Process Manager.

Keywords: Service Oriented Architecture, exoplanets discovery, distributed computing, service oriented computing, composition, choreography and orchestration, quality of service, calculations on graphics processors.

1 Introduction

We cannot understand Earth properly unless we can see it as a whole system; changing perspective, going out into space and comparing it to other planets is a good way to do it. Exoplanets are planets being discovered around stars other then our Sun. The recent year has been a great year for planets discovery, hundreds of them have been discovered in remote places and the range and variety of planets around other stars is much broader than anything we can see in our solar system. Every year we discover planets we wouldn't suspect could exist. However, we are still searching for Earth-like planets. Current news on European Space Agency project, Gaia [1, 2, 3] promises a three-dimensional map of our Galaxy, the Milky Way, revealing the composition, formation and evolution of the Galaxy. Gaia will provide unprecedented measurements of about one billion stars in our Galaxy and throughout the Local Group. As the project leaders claim, additional scientific products of the project include detection and orbital classification of tens of thousands of extrasolar planetary systems.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_66, © Springer International Publishing Switzerland 2014

The abovementioned facts, combined with the partnership with Project Solaris team ([4]), create a great opportunity to harness the new data and compare it with the exoplanets identification methods used in the Project Solaris [5, 6], which is one of five projects selected, out of 2200 in total, in celebration of five years of existence of European Research Council, making it one of most successful young Polish projects.

The constant development of technologies to support astronomers in their day-to-day work, allowed the Solaris Project researchers to implement the concept of autonomous robotic telescopes. Construction of robotic ground-based telescopes (and even entire observatories) has become one of the most important branches of instrumental astronomy. By a robotic telescope we understand a unit able to carry out at least one night of observations by itself, based on the general guidelines set by the user. Based on experience, however, such devices are able to work independently for months. This is a very significant departure from the traditional observational astronomy, in which the observer directly supervises the devices during all clear nights a year (in good observatories even ~ 300 nights). This means considerable relief from a very busy and costly obligation to conduct even a few weeks of nightly observations in remote locations. Research costs of the traditional approach are even higher, considering that observation telescopes are spread over several continents (such as South America, Africa and Australia).

This is also a right moment for adapting new software architectures, namely Service Oriented Architectures [7] and their iterations based on events, to use them to compliment the next generation hardware and methods used in astronomical projects. The benefits of SOA adaptation range from process and data mediation via autonomous web services [8], automated composite web service creation and execution [9] to guaranteeing quality of those distributed compositions through contracts (SLA – Service Level Agreement, [10, 12, 17]) and quality oriented planning [11, 18, 19, 20]. Recently, a number of works applying the issues of resource allocation strategies [15], content-awareness [16] and communication with mobile infrastructure components [14] to the service oriented architecture was published. Our architecture utilizes some of these results.

With SOA those projects will achieve not only greater flexibility but also more openness, allowing rapid growth of new client applications harnessing the acquired data.

2 Grounds for SOA and GPU Processing Adoption

2.1 Typical SOA Application

Currently most popular kinds of software solutions are either mobile or web-based. Amongst the latter are service-oriented solutions with Service Oriented Architecture as a flagship for enterprise applications. Big businesses, like banking industry, were first to adopt those solutions out of need for integration of legacy applications. The mentioned integration didn’t come cheap; however, cost of not being able to provide new solutions and new technologies for clients could be much greater and equivalent to the end of the company.

Business solutions made a case of using SOA as an integration tool, rescuing companies that couldn't afford rewriting its software. However, this creates a distorted view on the SOA paradigm, which could also show how to build new, lighter systems in which services not only provide functionality of large monolithic software, but also can be autonomous software components, easily scalable and replaceable.

The goal of presented project is to focus on the lighter, more agile SOA solutions developed in Wrocław University of Technology. Those solutions are aimed not only at business partners but also to help the scientific community to create backbones for various projects typically not focused on the IT perspective, but which would greatly benefit from a scalable and distributed system. Also, scientists – in contrast to businessmen – are more open to innovative, non-commercial solutions, which come with certain risks, but allow for automation of system reconfiguration and adaptation.

2.2 Project Solaris Requirements

The location of the Project Solaris telescopes (three stations on the three continents on the southern hemisphere) allows for observations 24 hours a day. Telescopes generate about 80 TB of data per year (in the form of hundreds of thousands of digital photos of the sky). Reduction, analysis and management of such data amount to a non-trivial amount of processing. To effectively use this data in astronomical research it is necessary to use modern methods for data transport and processing.

Currently, project infrastructure consists of a computing system based on GPGPU chip in each location. Although originally the GPUs are designed to render graphics, their architecture is fully capable of conducting parallel calculations. Today, it is tested how GPGPU-based systems can help in an ever-wider range of issues related to the processing of huge amounts of data. Many research institutions, including Massachusetts Institute of Technology, University of Cambridge take advantage of programmable GPUs to accelerate the speed of calculation.

On the other hand, a programming interface for GPU is still being improved. Two major companies, NVIDIA and AMD, noticing the huge market demand and research on this type of solution, created a dedicated architecture designed to perform numerical calculations. NVIDIA CUDA and AMD ATI Stream often not only expand, but also even supersede computing architectures based on CPU.

Astronomical observations are the ultimate images of the sky. The nature of these image data and of image reduction methods fit perfectly into the stream model offered by GPU computing units. The initial reduction in the preparation of observational data and further proper analysis, leading to scientific interpretation seem to be ideal to be approached with the map-reduce graphic units architecture. However, today's implementations of reduction algorithms usually do not fit the data being reducible. The common use of sequential programs for CPUs blocks the performance increase in two ways. First of all CPUs are not suitable for image processing (hence the original distinctions between CPUs and GPUs). Secondly, sequential characteristic of the method implementations means that programs are insensitive to distributed solutions (like SOA and cloud computing) and as such are a crucial bottleneck when processing vast amounts of observational data.

2.3 Indicators for SOA Adoption

The moment of observation and image reduction are clearly separated. Although the observations are limited to geographic locations, the analysis of observations does not have such requirements. So it is valid to consider that the reduction phase can exists as a service that uses the observation data from telescopes, but offers data ready for scientific interpretation. In addition, if we assume that this architecture uses the SOA paradigm, the physical location is hidden even from the reduction service, and the concept of distance is defined on the basis of QoS (Quality of Service).

For Project Solaris this means that computational units can be deployed in both the locations of observation, as well as in separate data centers, and the decision about the distribution of data for processing is carried out in the SOA platform. This provides a mechanism that is naturally ready to be extended to any number of computational units, capable of task relocation depending on data transportation costs, and prepared for analysis of observational data in an efficient way due to the GPUs, which can be installed in computation units, making the solution more cost efficient.

The great advantage of a dedicated computing architecture based on the SOA paradigm is that the computational units, combined with specific algorithms implementations, can be offered as a service. Currently, standards for data storage in astronomy are established. The same can be said for algorithms implementations and their input and output characteristics. Thus, more broadly, such a system can serve not only to a specific project and specific processing methods, but also to accelerate the results generation in astronomical projects in general.

Innovative SOA approaches, like service composition methods, allow for greater flexibility of presented solutions. Observation and processing requests can be transformed to composite services that create a unique, distributed workflow of processing algorithms. Such approach could be applied not only to exoplanets discovery but other astronomical task, as well as molecular biology, physics, market analysis and others. The fundamental requirements for adaptation of this solution are semantic description of the domain and its algorithms and, for the utilization of the GPU processing power, ability to perform map-reduce operations on the gathered data, so that it could be processed in uncorrelated, independent parts.

3 Platform Description

3.1 Platform Overview

Based on the analysis of our partners' requirements and current technology, including the results of recent research conducted at the Wrocław University of Technology, the proposed model of a distributed platform is based on Service-Oriented Architecture (SOA) extended with the concepts of event-driven architecture (EDA - Event Driven Architecture) and the use of knowledge management services (SOKU - Service Oriented knowledge Utilities).

The main objective of the proposed SOA model is to provide an infrastructure that will enable an easy and flexible management of all service-based applications running under the control of the platform and easy expansion of the platform with new functionality (provided by applications built in accordance with the proposed model, or external applications that provide their functionality via web services). This platform will also make available the possibility of redefining the business processes carried out by the Process Manager, as well as manual and automatic composition of composite services executed in the platform.

Applying service-based solutions (Service-Oriented Architecture) to astronomical research is a natural consequence of the requirements analysis for this area and led us to design a platform that will be able to:

- store, transport (on long distances) and process large amounts of data,
- (on the basis of a preliminary analysis of the data) share data and process them in parallel (also in a decentralized way, in locations relatively distant to each other),
- separate the requirements, control, data, processing and visualization layers,
- replicate data and autonomously, transparently manage their synchronization,
- virtualize and multiply stateless computing services,
- be extended with new computational methods provided via web services,
- adjust the allocation of computing resources, depending on the requirements,
- select and compose appropriate methods of processing, depending on the requirements.

In Figure 1 you can see a scenario, where a user defines his requirements using the appropriate web application of the integration platform (1). The presented integration platform allows for coexistence of various web applications with different graphical interfaces, designed for a specific user group or a specific task. Here, an astronomer will define, for example, when and which fragment of the sky he would like to observe using the projects infrastructure and then what scientific questions he would like to be answered. Those questions, through the use of composition methods (semantic-based, AI Planning, QoS-optimizing), will be then translated to a workflow of processing algorithms that will process the image and analyze the data – in our basic scenario allowing for extrasolar planets discovery.

The astronomer could define other requirements, like observing particular objects and not pieces of the sky, but all those requirements would be transferred to the Process Manager (2), where a plan for the complete process would be prepared, starting from schedule of observations, preprocessing of the data, and finally processing in the Computational Grid. This plan should take into consideration information about both: other requests from scientists using the platform and the platforms' current state (load in the network, telescopes, datacenters and computational grids). Using a centralized observation scheduler and composition services (3), the Process Manager can optimize the plan for each of the steps, predicting their estimated time and costs and selecting such a plan that meets the user non-functional requirements.

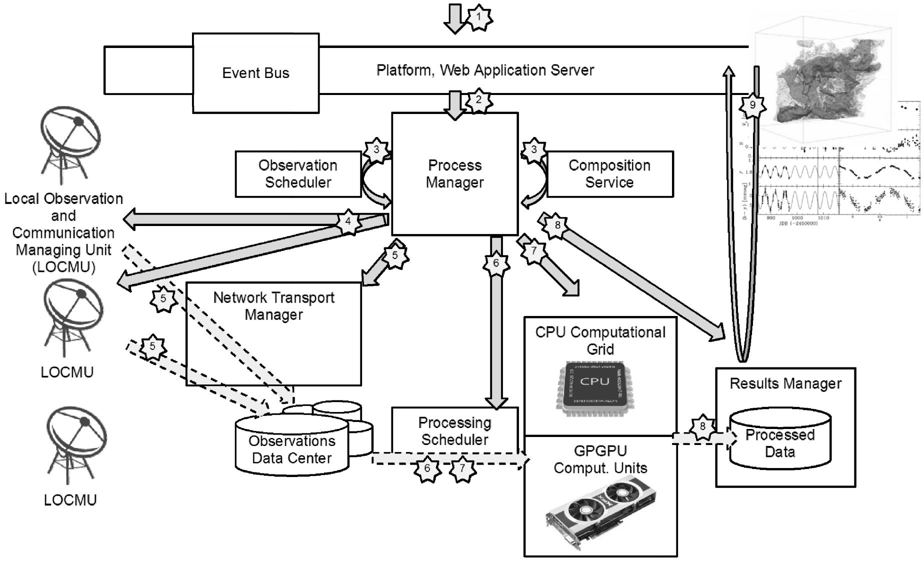


Fig. 1. A platform overview in an example scenario

When the plan is ready, it is executed. This is, in fact, a long-running business process that starts with transferring the observation schedule to selected telescopes (4). It could be one or more devices, depending whether they complete each other in an overlapping observation task (in the specified timeframe when a fragment of the sky stops being visible for one telescope it could be visible for another).

When the observation is concluded, observation data can be transferred to the Observations Data Center. However, it is about 4 GB of raw data that has to be transferred though half of the globe. To minimize the time and cost data reduction algorithms take place, cutting down the size of the data to only hundreds of MB. Then, the Process Manager is informed, via an event sent to the Event Bus, that the observation data is ready to be transferred.

Please, note that the data is not sent automatically, but the whole network transportation process is enveloped in a service managed by a highly specialized software component, the Network Transport Manager, which is also responsible for the quality issues and builds on the solutions previously developed for service systems in New Generation Networks [13]. It all begins, however, with an event. This is the first time an event-based communication has been mentioned, but the Event Bus accompanies the process at every step of its execution. It could not be deduced from Figure 2 simply because too many connections would have to be shown, practically obscuring the idea behind the platform. However, already during the observation gathering process in the telescope, several background events have been passed through to the Event Bus and more would be routed this way, had some errors occurred, requiring rescheduling and recomposition of the whole process.

The Process Manager, being informed by the event that it can proceed with the next step of the process execution, invokes the Network Transport Manager Service

(5), executing the predicted, partially reserved and prepared network transfer of the data. The role of the transport component is to guarantee that the data is transferred correctly through the web, if this is physically possible. The Network Transport Manager continuously monitors the network, especially the route pre-planned for the transfer and, knowing its limits and required data load, prepares appropriate transfer protocols and corrective mechanisms – all in cooperation with appropriate telescopes and Observations Data Center interfaces. When the transfer to the data center is finished, an event is sent via Event Bus to the Process Monitor.

The location of the data in the distributed Observations Data Center has been predefined during the initial planning. However, knowing where the data will be processed could be used to further optimize the location. It would be wasteful to calculate this precisely during user requirement analysis. That step had to be quick and the platform responsive, calculating only secure estimates. Still, while gathering the observations, a replanning operation could take place. Provided that the Observations Data Center is also an autonomous software component it can respond to the Network Transport Manager with new transfer coordinates when prompted, without having to burden the Process Manager (a notification event is sent regardless).

The schedule for the data processing has been partially prepared during the user request analysis and, similarly to the Observations Data Center location, it can be further optimized during the first steps of the process execution (6): observation gathering, transportation and synchronization. The main difference to this step is that it is expected that the user could change his mind about what kind of research and thus the kind of processing he requires from the system. To this point many reduction algorithms are lossless and even if this would not be the case, the number of applicable processing requests is still vast. During the process of observation and data transfer, the user can still independently interact with various web applications on the platform and change his mind and, consequently, the processing schedule. In general, this is the point, which the user can come back to during his further research, and request other processing services on the already gathered data, stored in the Observation Data Center.

After an optional recomposition of the processing service and rescheduling it to new computational units, the data is sent to the Computational Grid (7). In most cases, to obtain the highest performance, the data will be processed on the GPGPU grids. Both typical classification or identification tasks and especially image processing tasks are fit for parallel processing on graphics processors; however, typical astronomical algorithms have to be adapted for that purpose. As a backup, a cheaper infrastructure of PCs (equipped both with CPU and GPU but not in specialized GPGPU racks) is ready for non-priority calculations.

Final data, prepared for visualization, is transferred to appropriate result data centers (8), where it is semantically described and optimized for reading purposes.

With a final notification event sent to the user and appropriate web applications, the process is finished. Through all the steps, the Process Manager has been controlling and monitoring the process execution and all functionality has been delivered to it as a service.

Next, independently to the Process Manager, after receiving an appropriate notification (an email with a link or an in-app message), the user requesting the processing, can view its results in one or many specialized applications (9). Applications prepared for this purpose are web-based, mostly interactive with touch screen compatibility. User requests, results and observation data are stored in various integrated web applications of the platform and in appropriate components – usually Observation Data Center and Results Manager.

3.2 Physical Architecture

Software components described in previous sections are autonomous and highly optimized for the planned tasks. They will be installed on separate servers and integrated via web services, most of which will be invoked only by the Process Manager.

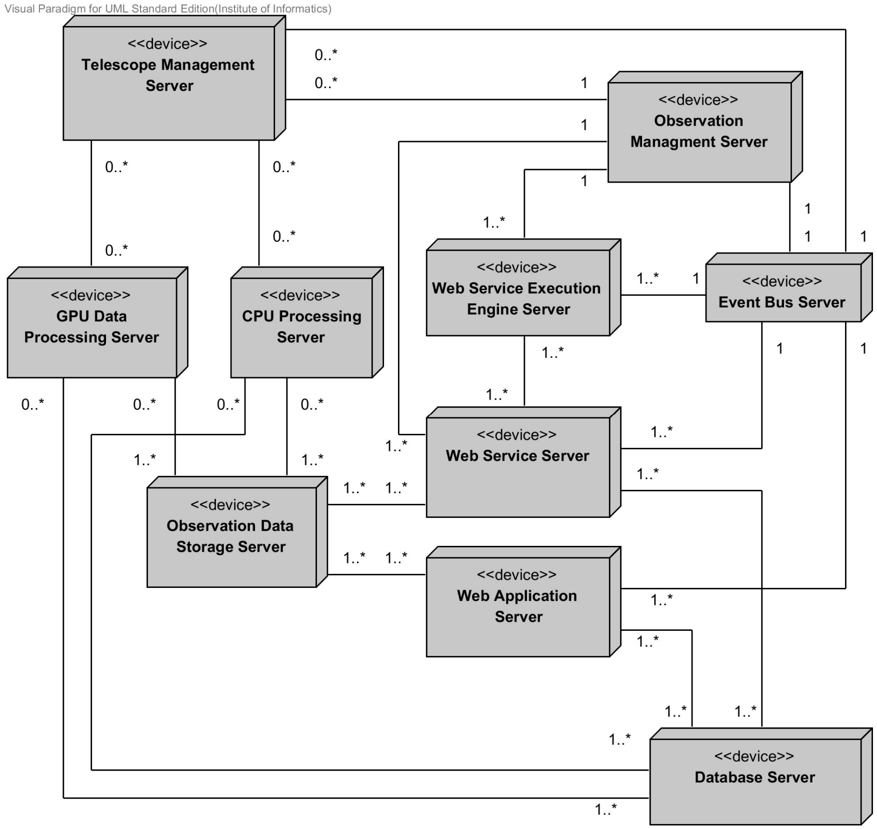

Visual Paradigm for UML Standard Edition(Institute of Informatics)
<<device>> Telescope Management Server
<<device>> Observation Managment Server
<<device>> GPU Data Processing Server
<<device>> CPU Processing Server
<<device>> Web Service Execution Engine Server
<<device>> Event Bus Server
<<device>> Web Service Server
<<device>> Observation Data Storage Server
<<device>> Web Application Server
<<device>> Database Server
0..*
1
1..*

Fig. 2. The physical architecture of the platform

Each component will have different load characteristics. The Process Manager will be executing multiple long lasting processes. Their nature allows for potential distribution of the process instances to separate servers.

The Event Bus will potentially route thousand of events per hour. Its distribution could be problematic, considering that some events are related and should be interpreted in context of other events.

The Computational Grid is distributed by default and after scheduling (and map-reduce approach etc.) each unit will be responsible for one task at a time.

The Data Base component will synchronize data among distributed Data Base Servers, considering best location for the data, minimizing their copies.

Specialized services like scheduling services and composition services are stateless calculation services and can be stored in a scalable cloud infrastructure. Each server of the cloud will be a specialized web service server.

The platform front end consists of numerous integrated web applications. Each web application can be stored on a separate server; however, apart from our applications, other external applications can be integrated with the system, each with its own storage solution. None specific solution will be enforced as long as applications abide by the platforms standards.

4 Conclusions and Further Work

In this paper architecture and its motivation for the SOA-based observation and computation infrastructure for extrasolar planets discovery has been discussed. Such distributed computing architecture, based on GPGPU units with dedicated reduction algorithms implementations is currently a modern, flexible solution, which is ready for the competitive provisioning of scientific results. Computing power is accessible via service calls, which hide the GPU programming issues on the user level.

The platform, utilizing methods (service composition, observation data processing, GPU parallel processing) and tools (SOA integration platform, composite service execution engine, autonomous robotic telescopes, GPU processing racks) created by the partners (Nicolaus Copernicus Astronomical Center in Warsaw/Toruń, Vratis Ltd. and Wrocław University of Technology) will be implementing the presented scenario. It is our future goal to further test those solutions and report on implementation results. SOA paradigm allows also to test and implement effective strategies for data sharing (via scientific-oriented Web portals) and experiment planning thus delivering an innovative and effective tools for the scientific community.

Acknowledgments. The research presented in this paper has been co-financed by the European Union as part of the European Social Fund.

References

1. `http://www.rssd.esa.int/index.php?project=GAIA&page=index`
2. Clark, S.: EJR-Quartz: Gaia – ESA's Galactic Census (2012) ISBN: 978-92-9221-043-4, ISSN: 0250-1589
3. Fletcher, K., McCaughrean, M.: ESA's Report to the 39th COSPAR Meeting (June 2012) ISBN: 978-92-9221-421-0, ISSN: 0379-6566

4. http://www.projectsolaris.eu/
5. Udry, S., et al.: Planets in multiple-star systems: properties and detections. In: RevMexAA (Serie de Conferencias), vol. 21, pp. 207–214 (2004)
6. Sybilski, P., Konacki, M., Kozłowski, S.: Detecting circumbinary planets using eclipse timing of binary stars–numerical simulations. Monthly Notices of the Royal Astronomical Society 405(1), 657–665 (2010)
7. SOA Reference Model Technical Committee. A Reference Model for Service Oriented Architecture, OASIS (2006)
8. Wu, Z., Ranabahu, A., Gomadam, K., Sheth, A.P., Miller, J.A.: Automatic Composition of Semantic Web Services using Process and Data Mediation. In: Proceedings of the 9th Intl. Conf. on Enterprise Information Systems (2007)
9. Pathak, J., Lutz, S.B.R., Honavar, V.: MoSCoE: A Framework for Modeling Web Service Composition and Execution. In: 22nd International Conference on Data Engineering Workshops. IEEE Computer Society (2006)
10. Anderson, S., Grau, A., Hughes, C.: Specification and satisfaction of SLAs in service oriented architectures. In: 5th Annual DIRC Research Conference, pp. 141–150 (2005)
11. Ko, J.M., Kim, C.O., Kwon, I.-H.: Quality-of-service oriented web service composition algorithm and planning architecture. Journal of Systems and Software 81(11), 2079–2090 (2008)
12. Świątek, P., Stelmach, P., Juszczyszyn, K., Prusiewicz, A.: Service Composition in Knowledge-based SOA Systems. New Generation Computing 30(2), 165–188 (2012)
13. Świątek, P., Rygielski, P.: Universal comunication platform for qos-aware delivery of complex services. In: Proceedings of the VIth International Scientific and Technical Conference, pp. 136–139. Publishing House Vezha&Co. (2011)
14. Świątek, P., Klukowski, P., Brzostowski, K., Drapała, J.: Application of wearable smart system to support physical activity. In: Advances in Knowledge-based and Intelligent Information and Engineering Systems, pp. 1418–1427. IOS Press (2012)
15. Grzech, A., Świątek, P., Rygielski, P.: Dynamic resources allocation for delivery of personalized services. In: Cellary, W., Estevez, E. (eds.) 13E. IFIP AICT, vol. 341, pp. 17–28. Springer, Heidelberg (2010)
16. Świątek, P., Juszczyszyn, K., Brzostowski, K., Drapała, J., Grzech, A.: Supporting Content, Context and User Awareness in Future Internet Applications. In: Álvarez, F., et al. (eds.) FIA 2012. LNCS, vol. 7281, pp. 154–165. Springer, Heidelberg (2012)
17. Grzech, A., Rygielski, P., Świątek, P.: Translations of service level agreement in systems based on service-oriented architectures. Cybernetics and Systems 41(8), 610–627 (2010)
18. Rygielski, P., Świątek, P.: Graph-fold: an efficient method for complex service execution plan optimization. Systems Science 36(3), 25–32 (2010)
19. Grzech, A., Świątek, P.: The influence of load prediction methods on the quality of service of connections in the multiprocessor environment. Systems Science 35(3), 7–14 (2009)
20. Grzech, A., Świątek, P.: Modeling and optimization of complex services in service-based systems. Cybernetics and Systems 40(8), 706–723 (2009)

Decision Making in Security Level Evaluation Process of Service-Based Applications in Future Internet Architecture

Grzegorz Kołaczek, Krzysztof Juszczyszyn,
Paweł Świątek, and Adam Grzech

Institute of Computer Science, Wrocław University of Technology, Wyb.Wyspiańskiego 27,
50-370 Wroclaw, Poland
{Grzegorz.Kolaczek,Krzysztof.Juszczyszyn,
Pawel.Swiatek,Adam.Grzech}@pwr.wroc.pl

Abstract. A method of decision making in security level estimation process of service-based applications in Future Internet architecture is proposed. We demonstrate how distributed services can be composed to form an application run within the Next Generation Network (NGN) infrastructure and their security level may be assessed. Our approach is illustrated by the experiments carried on exemplary application (virtual laboratory Online Lab, using Future Internet IPv6 QoS architecture), in which our method was evaluated against two types of attacks observed with the use of traffic anomaly detection methods.

Keywords: Anomaly Detection, Security Level Evaluation, Future Internet, Service Oriented Architecture.

1 Introduction

Service Oriented Architecture (SOA) is software paradigm that enables organizations to build, deploy and integrate services independent of the framework on which they are run [7]. The main idea about this architecture is that businesses which exploit SOA paradigm can respond faster to market opportunities and get more value from their existing technology assets [7]. The composition of Web services allows building complex workflows and applications [33],[36]. Besides the obvious software and message compatibility issues a good service composition should be done with respect to the Quality of Service (QoS) (esp.: security) requirements [15-16].

The security evaluation process should be based on some formal prerequisites. The first problem is that the security measure does not have any specific unit. Also, security level has no objective grounding but it only in some way reflects the degree in which our expectation about security agree with reality; security level evaluation is not fully empirical process.

Security issues become crucial if we assume that complex processes are being realized by workflows of atomic services, which may have different security levels. The composition of Web services allows building complex workflows and applications on

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_67, © Springer International Publishing Switzerland 2014

the top of the SOA model, with preserving non-functional requirements if necessary [18],[19],[20]. In our work we address a Future Internet architecture in which a service-based application is being checked in order to assess the security level of its component services.

The main contribution of this work is a proposal of a novel approach for the decision making in the security level estimation process of service-based applications using a Future Internet architecture IPv6 QoS. We give also a brief overview of the Online Lab application which is a virtual computational laboratory and serves as an example and source of test data for the evaluation of our approach in section 2. Section 3 introduces security issues in the context of service architectures, while section 4 presents experimental results basing on traffic data gathered during normal operation of Online Lab and generated attack data. In the last section we conclude our work and point out directions of future works.

2 Systems Architecture

In this work we consider an IPv6 QoS system architecture [29],[30] developed in the polish national project IIP (polish acronym for Future Internet Engineering)[25]. In this architecture it is assumed that the system consists of multiple layers each of which provides certain functionalities to the adjacent upper layer. The first layer is a physical network infrastructure which with use of virtualization techniques provides to the second layer virtualized networking environment with dedicated communication and processing resources [25]. Such virtualization allows for coexistence of multiple isolated virtual networks (called parallel internets - PI), characterized among others by different frame formats, protocol stacks and forwarding capabilities, in a single physical infrastructure.

IPv6 QoS system is one of parallel internets existing in a virtual networking environment. In general the IPv6 QoS architecture is based on coupling of the Differentiated Services (DiffServ) [24] quality of service assurance model and Next Generation Network (NGN) signaling system. DiffServ is responsible for delivery to traffic flows generated by users required level of the quality of services by means of flow admission control, classification of flows to predefined traffic classes and processing of aggregated flows from different traffic classes [28]. The NGN signaling system is used to provide end-to-end QoS guaranties by reserving necessary amount of communication resources to each particular connection request. Reservation of communication resources is performed by assignment of the request to proper DiffServ traffic class, which meets the QoS requirements for this flow.

The purpose of signaling in NGN is twofold. The first one is to reserve required communication resources and to establish an end-to-end connection between a pair of hosts in the system. This signaling is performed at the network layer in so-called transport stratum. Second type of signaling is performed at the application layer (service stratum). Service stratum signaling is in general an application specific signaling (e.g. SIP signaling) the aim of which is to configure distributed modules of an application and to process information necessary to send to transport stratum a request for communication resources reservation. Signaling can be also viewed as a middleware which separates the networking layer functionalities and application domain-specific specific functionalities.

2.1 Virtual Computational Laboratory Online Lab

Virtual laboratory infrastructure should automate most tasks related to the reliable and reproducible execution of required computations [21],[23]. The application Online Lab is a distributed, service-based computational laboratory benefiting from the IPv6 QoS architecture which is used to distribute computational tasks while maintaining the quality of service and user experience [22]. Online Lab functionality allows definition, storing, sharing and execution of code and data. The communication mechanisms are benefiting from the Future Internet communication architecture and are designed for optimization of the users' Quality of Experience, measured by the response delay. Online Lab allows its users, i.e. students or researchers, to access different kinds of mathematical software via Python math libraries and perform computations on the provided hardware, without the need for installing and configuring any software on a local computer. The communication mechanisms are designed for optimization of the users' Quality of Experience, measured by the response delay. The functionality of Online Lab embraces:

(i) access to computational services ensured by user's virtual desktop which is windowed interface opened in a Web browser,

(ii) creation and removal of computational services with no limitations being assumed on the nature of computations – the users may freely program computational tasks in any language interpreted by running computational services,

(iii) user profile maintenance and analysis – the users are distinguished by their profiles which hold information about their typical tasks and resource consumption.

Online Lab (OL) implements an architecture consisting of user interface (OL-UI), core server (OL-Core), services and computational engines (OL-Services, based on the Python engine equipped with the set of math libraries in the current prototype). OL-UI is a web service emulating a desktop and a window manager. Code is being typed into specialized data spaces - notebooks, which are executable documents executed by OL-Services. A notebook is a collection of cells of different kinds, that allows for storing rich contents (text, tables, images, LaTeX rendered mathematics) and source code, and allow for evaluating this source code on remote machines that are equipped with required numerical and mathematical software. In addition the users may attach external data files to the notebooks and refer to them in the code cells. The idea is that users are required to have just a modern browser, like Firefox or Chrome (which run OL-UI), installed on their computers and this is sufficient to be able to perform high-end computing.

The process of user's query execution is presented in Fig. 1. OL-Core and OL-Services belong to the service. One notebook represents one computational task. The system also may recommend notebooks of other users. The content of the notebooks is annotated with the help of domain (Math) ontology.

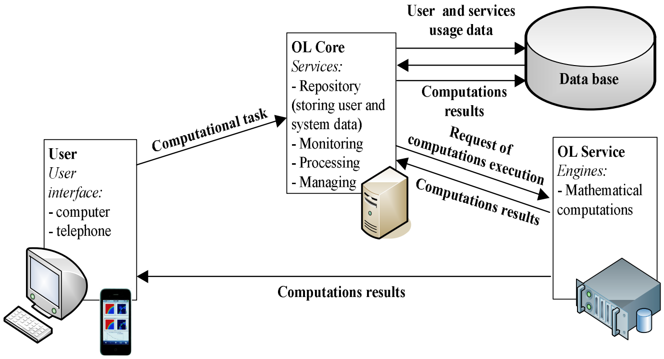


Fig. 1. The general schema of the Online Lab service execution

Additionally, OL-Core is constantly monitoring the OL-Services, storing execution times and data transfer volumes in its database. On the basis of first test implementation of Online Lab we have proposed the classification of computational tasks with respect to the data volumes and CPU time needed to complete the task:

(i) Normal (N) tasks,
(ii) Computation-intensive (CI) tasks,
(iii) Data-intensive (DI) tasks,
(iv) Data-Communication (DCI) intensive tasks.

An additional unique feature of Online Lab is the possibility of implementing dedicated computational services which may be available to other applications (in the firsts tests of this approach a dedicated Online Lab service was used as an analysis data module for an IPv6 QoS E-health application (patient monitoring).

3 Security in Service Oriented Architecture

The final success of the SOA concept and Future Internet applications can be obtained if many groups, both internal and external to the organization, contribute to the execution of a business process. Because in most cases the most valuable and also sensible part of each organization is information, a business partner is much more willing to share information and data assets, if it knows that these assets will be protected and their integrity maintained. Business partners will also be more likely to use a service or process from another group if it has assurance of that asset's integrity and security, as well as reliability and performance. Therefore ensuring security is a one of the most crucial elements while putting SOA approach into practice. Security issues become crucial if we assume that complex processes are being composed of atomic services which may have different security properties.

In this context, estimating the security of services on demand (i.e. when the composite service is composed) by existing methods and tools may be impossible or provide insufficient information. In described method a multi-agent approach has been proposed to evaluate security of atomic and composed services. Multi-agent technology can reduce the bandwidth requirement and tolerate the network faults - able to operate without an active connection between clients and server. As the security evaluation process must be accurate and efficient, these basic features relevant to agent and multiagent systems are the main motivation for many researchers to apply multiagent approach to the tasks related to system security. The second premise in this case is the correspondence of the multiagent environment to SOA systems. Multiagent systems are composed from the number of autonomous and mobile entities that are able to act both cooperatively and separately. The fundamental concept for SOA system is service – entity that could be evoked individually as well as in cooperation with other services. And at last, both multiagent and SOA systems tend to act in heterogenic and highly distributed environment.

When using multiagent system, agents can gather knowledge continuously and not only about security at one time but about its estimate in time as well. This typical for agent systems asynchronous information gathering and aggregation is more natural to the area of the problem of composite service security estimation. And at last, both multiagent and SOA systems tend to act in heterogenic and highly distributed environment.

As the number of SOA implementation grows the concerns about SOA systems security also increases. The literature related to the security of SOA focuses on problems with threat assessment, techniques and functions for authentication, encryption, or verification of services [1],[2],[6]. Some other works focus on high level modeling processes for engineering secure SOA [4],[9] with trust modeling [7], identity management and access contro [10,[12]. Many studies focus on secure software design practices for SOA, with special interest in architectural or engineering methodologies as the means to create secure services [3],[5].

In most cases building composite services converts into a constraint satisfaction problem - there can be many candidates (atomic services) for building blocks of a complex service (process) an and it is necessary to select the optimal execution plan. The required composition is expected to satisfy chosen QoS parameters [17]. Several approaches to the assessment of QoS parameters of composed services have been proposed so far but the first approach to the estimation of security level of a composite service was presented in [11]. The solution proposed in this work extends this approach by presenting a framework for practical opinion assessment of the agents, based on the first experiences with its practical implementation.

3.1 Multiagent Framework for Composite Services Security Evaluation

A composite service is a service that consists of several atomic services. Each of those services can be an independent application from a different service provider. In the construction of appropriate tool for monitoring and controlling all components of a Service Oriented System the following assumptions are valid:

- data must be acquired periodically in order to be up-to-date, due to the distributed nature of the system (and its typically large size) it is impossible to establish a centralized measuring system.

The decomposition of the measuring system raises questions about how the measured data should be aggregated, especially when taking into consideration the heterogenic nature of the information from various agents. These aspects were taken into account in the design of multiagent system architecture.

This section presents the architecture of composite services security evaluation framework. The adopted approach extends the security evaluation approach presented in [11]. The main element of the multiagent system architecture for SOA security evaluation is definition of the agent classes. The following agents classes have been considered:

- ASL – (Agent for Service Layer) agent evaluating the security of services (the name corresponds to the layered SOA security architecture presented in [4], in the simplest case we may consider only one layer of agents responsible for all atomic services).
- AM – Managing Agent

In this work we propose a two-layer agent framework for security assessment of composite service. In the first layer ASL agents periodically evaluate the security of atomic services. ASL agents can evaluate the security of many services and for each manifest different behavior (frequency of estimations, precision of tests). Sets of atomic services observed by ASL agents can mutually overlap as shown on Fig 2.

In the second layer in order to estimate the security level of a composite service a managing agent AM can gather security level estimates from corresponding ASL agents. Based on their opinions about atomic services and agents' trust level the AM can aggregate the opinions to build an overall security level estimate about the composite service. The managing agent AM performs multistage security estimation in response to the composite service security level estimation request. First of all the Managing Agent gathers appropriate opinions on services (generated by various ASL agents) from the history.

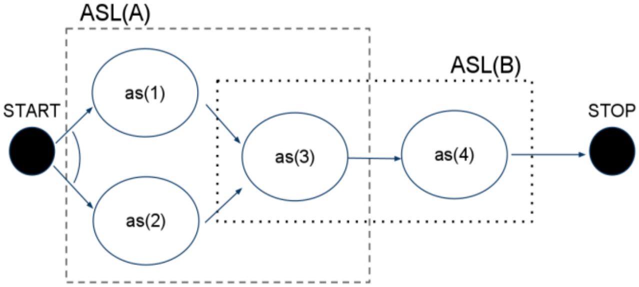


Fig. 2. An example of a composite service with overlapping observation areas

Various ASL agents can give many security level estimates of an atomic service (i.e. repeat evaluation in some time intervals). However, methods for aggregation of such security level estimate of any particular service are not described in this paper (will be developed in the course of future works). Here we assume that for each agent only one (up-to-date) security level estimate of an atomic service is available. In the next sections we present methods that allow for such security level estimation of each monitored atomic service and then methods for aggregation of such estimates of security level of atomic services into one security level estimate of a composite service.

4 ASL Implementation an Experimental Evaluation

The proposed ASL implementation for security level estimation of service-based applications in Future Internet uses anomaly detection approach. More precisely, the method benefits form time series analysis. The anomalous behavior of the systems is determined using the values for the behavioral attributes within a specific context. An observation might be an anomaly in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. Contextual anomalies have been most commonly explored in time-series data [19][20].

The implemented in ASL algorithm of anomaly detection in time series can be applied to detect anomalies in various types of values measured during the service execution time. The only requirement is that the values must constitute time series (e.g. memory and CPU usage level, a number of incoming and outgoing bytes, etc.). The detected anomalies are often related to various types of attacks. For example, high level of CPU utilization level or remarkably greater volume of data received by a service usually can be observed during denial of service (DoS) attacks. Other more specific types of attacks e.g. traffic injection attack also can be detected by time series anomaly detection methods. As this type of attack imposes extra processing effort of a service, it should be noticed during CPU utilization level analysis. Another example of attack, a ruffling attack disrupts user requests spacing and creates traffic bursts or abnormal interarrival times, what can be noticed at time series describing incoming or outgoing traffic and number of user requests. The examples of practical application of time series analysis to detect abnormal service behavior and possible attacks in Online Lab environment are presented and discussed further in this section.

The typical behavior of the most computer systems shows some periodicity, e.g. number of processes executed during the day time or data transferred. The length of the characteristic period varies from system to system but typically the most significant correlations in system parameters values can be noticed in a day and a week long periods. In our experiments we assumed that also data rate, volume, etc. characterizing services executed in Online Lab show this type of dependencies.

The general idea of detection algorithm implemented by ASL is as follows. First, time series is created (from the starting point of measurement x1 the current i-th element):

$$X_{T,i}=\{x_1, x_2, x_3,\ldots, x_j, x_{j+1},\ldots, x_i\} \quad (1)$$

where elements of $X_{T,i}$ are values of measured data volume transferred by the selected Online Lab service. Apart from $X_{T,i}$ time series two other families of sub-time series are analyzed by anomaly detection algorithm. The first one:

$$X_{S,i}=\{x_i, x_{i+P}, x_{i+2P},\ldots, x_{i+kP}\} \tag{2}$$

where elements of $X_{S,i}$ are values taken from $X_{T,i}$ and where each two subsequent elements are in a distance of *P*. *P* is a value describing period length, in our case it equals to 24 hours (1 working day).

The second family of sub-time series:

$$X_{L,i}=\{ x_i, x_{i+1}, x_{i+2},\ldots, x_{i+P-1}\} \tag{3}$$

where elements of $X_{L,i}$ are all subsequent values taken from time series $X_{T,i}$ from a particular *i-th* period of observation (Fig.3).

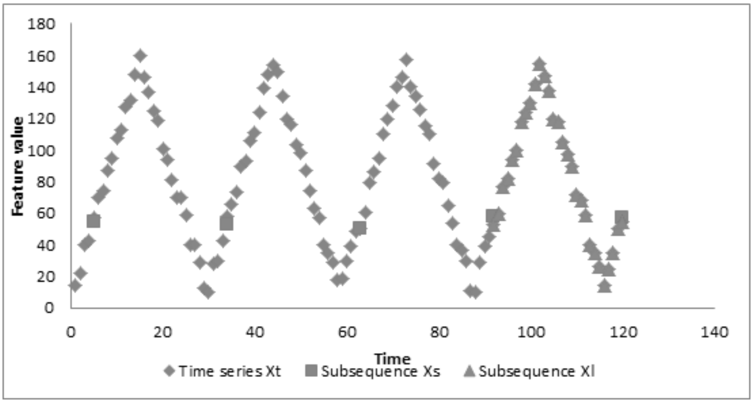


Fig. 3. Time series analysis for anomaly detection

For each one of the above described families of time series are evaluated the exponentially weighted moving average values using standard EMA (exponential moving average) formula:

$$\overline{X}_{T,i} = \overline{X}_{T,i-1} + w*(x_i - \overline{X}_{T,i-1}) \tag{4}$$

where $\overline{X}_{T,i}$ is exponential moving average calculated for $X_{T,i}$ time series at i-th point and w is a coefficient with empirically assigned value. In the corresponding way the values of exponential moving average of $\overline{X}_{L,i}$ and $\overline{X}_{S,i}$ are calculated. The observed values characterizing behavior of Online Lab service are analyzed in three dimensional space (time series $X_{T,i}$, $X_{L,I}$, $X_{S,i}$). This multidimensional analysis improves the precision of anomaly detection.[11] Especially, taking three dimensions together allows for better understanding the seasonal and trends changes appearing in the time series.

For each time series the estimates of appropriate standard deviation ($\sigma_{T,i}$) and local difference ($\delta_{T,i}$) are evaluated in the following way:

$$\delta_{T,i} = \left| \overline{X}_{T,i} - x_i \right| \tag{5}$$

and

$$\sigma_{T,i} = \sqrt{\frac{1}{i}\sum_{j=1}^{i}(x_j - \overline{X}_{T,i})^2} \tag{6}$$

The $\sigma_{T,i}$ estimates the measure of variability of $X_{T,i}$ in time series values and $\delta_{T,i}$ evaluates how much the current observation differs from the average at the current time point i. The values $\sigma_{S,i}$, $\sigma_{L,i}$ and $\delta_{S,i}$, $\delta_{L,i}$ for two remaining time series are calculated in the correspondent way.

Using defined over here estimates of standard deviation we define the ASL's opinion ω = ⟨b, d, u⟩ about security level of Online Lab service which will be used for evaluation of the whole Online Lab application security level. The notion of the opinion and the corresponding formal model has been defined by Josang in [11-14].

The formal definition of disbelieve value in time series analysis during security level evaluation process is given by the following formula:

$$d = \min\left\{ \frac{1}{2\sqrt{3}} \sqrt{\left(\frac{\delta_{S,i}}{\sigma_{S,i}}\right)^2 + \left(\frac{\delta_{L,i}}{\sigma_{L,i}}\right)^2}, 1 \right\} \tag{7}$$

The disbelieve value d ranges from 0 to 1. When detected anomaly is relatively small (near the average values) the d value will be near 0. While we observe the high deviation from the earlier observed values (three times greater than standard deviation) the disbelieve value d equals to 1.

The uncertainty value u in opinion ω = ⟨b, d, u⟩ about security level of the monitored communication link is evaluated using the following formula

$$u = \begin{cases} 0 \quad if \quad d = 1 \\ \min\left\{(1-d), \frac{\sigma_{S,i}}{\sigma_{L,i}}\right\} \quad if \quad d \neq 1 \end{cases} \tag{9}$$

where $\frac{\sigma_{S,i}}{\sigma_{L,i}}$ denotes the proportion between estimates of variance calculated for the last period of observation and the variance calculated for all observations from $X_{S,i}$ sub-time series.

4.1 Experiment Description

In our experiment the traffic between component services of the Future Internet application presented in earlier section is being observed in order assess its security. In an experimental setup the Online Lab underwent a 42 hour-long test involving the user group consisted of 38 users – two student groups (2x16 persons) and 6 independent researchers.

Students participated in an online course, with requirement to complete their tasks in 48 hours while the Researchers were free to use the application. The main goal was to evaluate the logging system and analysis methods which were to return the statistics allowing to properly reserve resources required for Online Lab, and –during next phase – predict the resource consumption assuming that the user group is known and observed for some time. The results fall into two main categories – performance evaluation and user action analysis. Online Lab performance evaluation

For the period under consideration the parameters of OL-Services and all communication links were measured and stored. For example we have noticed, that most of the elementary actions involving the evaluations of worksheet cells require times below 1 second (the problems solved were not complex in this experiment) – Fig. 4. However, there were exceptions – some operations involved significantly larger data transfer and computational time.

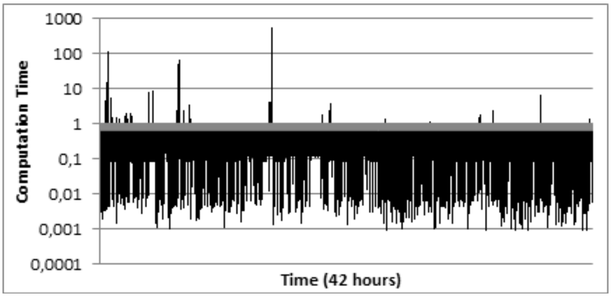


Fig. 4. Time needed for the evaluation of notebook cell (~4200 events)

The above statistics, stored and related to the worksheets content via the Online Lab worksheet annotation system are intended to be used in order to estimate the need for OL-Services and parameters of communication links between the services of the Online Lab, thus providing of content- and context awareness capability of the Online Lab, within the model presented in [21],[31]

We have recorded the actual traffic volumes occurring between the services of the Online Lab application which was in our experiment added to the pattern characteristic for chosen types of the attacks. The resulting traffic logs were used to apply our method in order to detect malicious traffic.

4.2 Security Level Estimation

The data obtained during a test phase have been used to demonstrate the described in earlier sections approach to security level estimation of services in SOA systems. The first test scenario investigates the feasibility to detect by ASL anomalies related to some specific security threats as traffic injection and traffic ruffling. In this scenario ASL analyzes a data stream characterizing data exchange among services in Online Lab. This characteristic includes information about number of bytes sent through each link between each pair of services. The corresponding data stream is provided to ASL through Online Lab monitoring data repository which collects data from the monitoring module.

In the first example of this scenario attacks have been simulated by disturbing the data volume transmitted by a service by injecting some malicious traffic. At the beginning, the typical traffic generated by OL-Services, OL-Load Balancer and OL-Core has been captured during test phase descripted in the earlier section of this paper. Next, the captured traffic has been reengineered to simulate traffic injection and ruffling attacks. Using tcpreplay tool the previously captured traffic has been resend 7 times (*tcpreplay --loop=7 --intf1=eth0 u1_u2.pcap*). Simultaneously, the ASL using algorithm described earlier in this section and data stream provided by Online Lab management created time series characterizing typical traffic volume of OL-Services. ASL recognizes the traffic characteristic as typical behavior and reports it as an event without risk opinion $\omega = \langle b=1, d=0, u=0 \rangle$. After some 2 intervals the malicious traffic has been injected. The additional packets have been generated by tcreplay tool (*tcpreplay --loop=30 --intf1=eth1 u1_u2.pcap*) and new values of time series describing traffic volume related to selected OL-Service have been analyzed by ASL. ASL detects that traffic volume of OL-Service has been changed and generates reports with corresponding value of opinion about OL-Service security level $\omega = \langle b<1, d>0, u>0 \rangle$. The plot of the changes in opinion values for each of the measurement intervals illustrates figure 5. As the injected traffic infers more time series elements the disbelief increases. After some time the additional packets, as they are generated with the constant distribution of packets' inter-departure time value, do not more increases disbelief values. The uncertainty value u grows (and decreases) in the corresponding way to the changes introduced by the additional traffic volume.

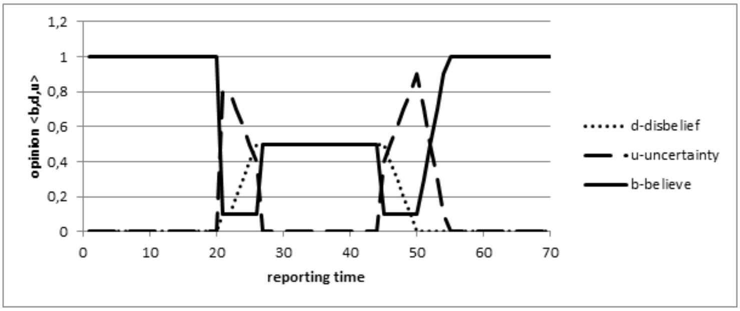


Fig. 5. Traffic injection attack

The next traffic related attack scenario shows the simulated traffic ruffling. The typical traffic generated by OL-Service has been simulated using the same tools and values of configuration parameters as in previous example. After some 2 intervals the attack starts and in a consequence the traffic between OL-Service and OL-Core has been disrupted. The ruffling attack has been simulated by modification of the packets generator parameters values to *tcpreplay --multiplier=5.2 --intf1=eth0 u1_u2.pcap* which means that the captured traffic has been replied 5.2 times faster than it was captured. The plot of the changes in severity and intensity values illustrates figure 6. This experiment shows that this type of attack generates more fluctuations in disbelief value than the traffic injection attack. It is the effect of the overlapping different periods of the typical and malicious traffic. This type of attack at communication links between Online Lab services infers the time series in more complicated way what can be seen in figure 6.

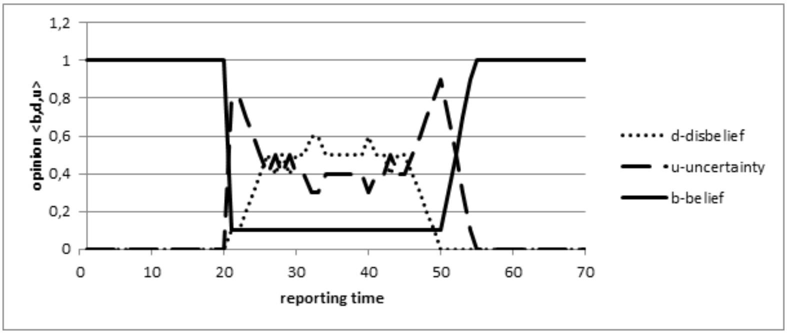


Fig. 6. Traffic ruffling attack

This difference can also be used to distinguish different type of attacks against OL-Services. The attack type recognition using collected from security evaluation module values of disbelief is the interesting aim for the further research.

Note that, when the opinions about the security level of all the services are assessed, we can use the operators introduced in [14] to build an opinion about service based-application as a whole.

5 Conclusions

We have presented a decision making process for security level estimation of service-based applications in Future Internet architecture, which utilizes opinions and the Subjective Logic's formal model. The proprieties of the chosen methods of traffic anomaly detection in the context of this approach has been illustrated with experiments on real traffic data to which the attack schemes were added. The data came from a service-based distributed application dedicated for the use in Future Internet

IPv6 QoS architecture. Our approach to decision making about security level is flexible and allows introduction of the anomaly and attack measures other then used in our examples (the only condition is the ability of expressing results in the form of Subjective Logic opinions). This remark shows the way for future experiments which will aim to develop autonomous agents acting within Future Internet architecture and being capable of applying different methods of security assessment to distributed, service-based applications.

References

1. CERT (2009), `http://www.cert.org` (retrieved March 20, 2009)
2. Epstein, J., Matsumoto, S., McGraw, G.: Software security and SOA. IEEE Security and Privacy 4(1), 80–83 (2006)
3. Fernandez, E.B., Delessy, N.: Using patterns to understand and compare web services security products and standards (2006)
4. Kolaczek, G.: Opracowanie koncepcji specyfikacji metod i modeli szacowania poziomu bezpieczeństwa systemów SOA i SOKU, WUT (2009) (in polish)
5. Nakamura, Y., Tatsubori, M., Imamura, T., Ono, K.: Model-driven security based on web services security architecture. In: IEEE International Conference on Services Computing, vol. 1, pp. 7–15 (2005)
6. SANS Institute (2006), `http://www.sans.org` (retrieved March 20, 2009)
7. Skalka, C., Wang, X.: Trust by verify: Authorization for web services. Paper presented in ACM Workshop on Secure Web Services, pp. 47–55 (2004)
8. SOA Reference Model Technical Committee. A Reference Model for Service Oriented Architecture, OASIS (2006)
9. Steel, C., Nagappan, R., Lai, R.: Core security patterns: Best practices and strategies for J2EE, web services, and identity management. Pearson, Upper Saddle River (2006)
10. Tari, Z., Bertok, P., Simic, D.: A dynamic label checking approach for information flow control in web services. International Journal of Web Services Research 3(1), 1–28 (2006)
11. Kolaczek, G., Juszczyszyn, K.: Smart Security Assessment of Composed Web Services. Cybernetics and Systems 41(1), 46–61 (2010)
12. Jøsang, A.: A Logic for Uncertain Probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–311 (2001)
13. Jøsang, A.: A Metric for Trusted Systems. In: Proceedings of the 21st National Security Conference, NSA, pp. 68–77 (1998)
14. Jøsang, A.: Conditional Inference in Subjective Logic. In: The Proceedings of the 6th International Conference on Information Fusion, Cairns, pp. 279–311 (2003)
15. Anderson, S., Grau, A., Hughes, C.: Specification and satisfaction of SLAs in service oriented architectures. In: 5th Annual DIRC Research Conference, pp. 141–150 (2005)
16. Milanovic, N., Malek, M.: Current Solutions for Web Service Composition. IEEE Internet Computing 8(6), 51–59 (2004)
17. Frolund, S., Koisten, J.: QML: A language for quality of service specification (1998), `http://www.hpl.hp.com/techreports/98/HPL-98-10.html`
18. Charif, Y., Sabouret, N.: An Overview of Semantic Web Services Composition Approaches. Electronic Notes in Theoretical Computer Science 146, 33–41 (2006)

19. Salvador, S., Chan, P.: Learning states and rules for time-series anomaly detection. Tech. rep., 2008 CS–2003–05, Department of Computer Science, Florida Institute of Technology Melbourne (2003)
20. Weigend, A.S., Mangeas, M., Srivastava, A.N.: Nonlinear gated experts for time-series: Discovering regimes and avoiding overfitting. Int. J. Neural Syst. 6(4), 373–399 (1995)
21. Noguez, J., Sucar, L.E.: A Semi-open Learning Environment for Virtual Laboratories. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 1185–1194. Springer, Heidelberg (2005)
22. Pautasso, C., Bausch, W., Alonso, G.: Autonomic Computing for Virtual Laboratories. In: Kohlas, J., Meyer, B., Schiper, A. (eds.) Dependable Systems. LNCS, vol. 4028, pp. 211–230. Springer, Heidelberg (2006)
23. Juszczyszyn, K., Paprocki, M., Prusiewicz, A., Sieniawski, L.: Personalization and content awareness in online lab – virtual computational laboratory. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS, vol. 6591, pp. 367–376. Springer, Heidelberg (2011)
24. Blake, S., et al.: An architecture for differentiated services. RFC2475 (1998)
25. Burakowski, W., et al.: The Future Internet Engineering Project in Poland: Goals and Achievements. In: Future Internet Poland Conference, Poznan, Poland (October 2011)
26. Mosharaf Kabir Chowdhury, N.M., Boutaba, R.: A survey of network virtualization. Computer Networks: The International Journal of Computer and Telecommunications Networking 54(5), 862–876 (2010)
27. Grzech, A., Rygielski, P., Świątek, P.: Translations of Service Level Agreement in Systems Based on Service-Oriented Architectures. Cyb. and Systems 41, 610–627 (2010)
28. ITU-T Rec. Y. Functional requirements and architecture of next generation networks (2012)
29. Tarasiuk, H., et al.: Provision of End-to-End QoS in Heterogeneous Multi-Domain Networks. Annals of Telecommunications 63(11) (2008)
30. Tarasiuk, H., et al.: Performance Evaluation of Signaling in the IP QoS System. Journal of Telecommunications and Information Technology 3, 12–20 (2011)
31. Rygielski, P., Tomczak, J.M.: Context Change Detection for Resource Allocation in Service-Oriented Systems. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS, vol. 6882, pp. 591–600. Springer, Heidelberg (2011)
32. Rygielski, P., Świątek, P.: Graph-fold: an efficient method for complex service execution plan optimization. Systems Science 36(3), 25–32 (2010)
33. Świątek, P., Stelmach, P., Prusiewicz, A., Juszczyszyn, K.: Service composition in knowledge-based SOA systems. New Generation Computing 30(2/3), 165–188 (2012)
34. Świątek, P., Rygielski, P., Juszczyszyn, K., Grzech, A.: User assignment and movement prediction in wireless networks. Cybernetics and Systems 43(4), 340–353 (2012)
35. Świątek, P., Juszczyszyn, K., Brzostowski, K., Drapała, J., Grzech, A.: Supporting content, context and user awareness in Future Internet applications. In: Álvarez, F., et al. (eds.) FIA 2012. LNCS, vol. 7281, pp. 154–165. Springer, Heidelberg (2012)
36. Fraś, M., Grzech, A., Juszczyszyn, K., Kołaczek, G., Kwiatkowski, J., Prusiewicz, A., Sobecki, J., Świątek, P., Wasilewski, A.: Smart Work Workbench: Integrated tool for IT services planning, management, execution and evaluation. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 557–571. Springer, Heidelberg (2011)

Integrating SIP with F-HMIPv6 to Enhance End-to-End QoS in Next Generation Networks

Muhammad Zubair[1], Xiangwei Kong[1], Irum Jamshed[2], and Muhammad Ali[2]

[1] School of Communication and Information Engineering,
Dalian University of Technology, Dalian, China
m.zubairpaf@gmail.com,
kongxw@dlut.edu.cn

[2] Center of Excellence in Information Technology,
Institute of Management Sciences, Peshawar, Pakistan
irum.jamshed@gmail.com,
muhammad.ali@imsciences.edu.pk

Abstract. With the advancement in communication technologies, user's demand continuously increases while moving across diverse networks. The research community is putting their best efforts in the deployment of Next Generation Networks (NGNs). The primary goal of NGNs is to provide Always Best Connected (ABC) services. Researchers' current focus is on many issues such as support for multicasting and QoS, security, resource management and allocation, location coordination and handoff. In such a heterogeneous environment, ensuring the end-to-end QoS is a challenge. The problem in combining Session Initiation Protocol (SIP) with Mobile IPv6 and Seamless MIPv6 is that the traffic goes through core network every time the call/ sessions are established, resulting into very high delay and overhead on core networks. In this paper, integrating SIP with Fast handover for Hierarchical Mobile IPv6 (ISF) is proposed which utilizes the balancing capabilities of each protocol and endeavors at dropping their practical dependencies. ISF ensures end-to-end QoS for multimedia session as well as minimizing the signaling traffic between edge and core networks, service disruption, and user authentication. The conducted results show that our proposed scheme outperforms the existing approaches in terms of handover latency, packet loss, and packet load.

Keywords: Fast Hierarchical Mobile IPV6, Next Generation Network, Session Initiation Protocol, End-to-end QoS.

1 Introduction

In the present age of communication, new mobile devices such as Apple's iPhone, Google's NexusOne, Android, Laptop and PDAs are becoming progressively popular and extensively used. This trend will increase in the near future. These devices contain multiple wireless network interfaces such as Bluetooth, Wi-Fi, WiMAX, Wireless Broadband (WBro), WLAN, 3G and 4G Mobile networks. Mobile users will

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_68, © Springer International Publishing Switzerland 2014

experience cost effective IP-based advanced real time applications such as VoIP, mobile games, mobile APTV, Emergency Telecom Services (ETS), Voice 2.0 and much more [1].

Over the past few years QoS and mobility in wireless networks has grasped researchers' attention. The growing demands and expectations of users for being connected are increasing due to which a lot of research is happening in this area. Wireless networks have evolved over last few years from the rudimentary level analog based First Generation (1G) networks to more efficient and digital Second and Third Generation (2G & 3G) networks. As a result, we are now stepping into Next Generation Networks (NGNs) also called Future Networks [2]. The goal of NGNs is to provide the users with Always Best Connected (ABC Concept) facility [3].

NGNs are based on "All IP based paradigm" that provide fully converged services, mobile access in an ambiguous manner, and support to heterogeneous devices [4]. Currently, two protocols are center of attention for supporting mobility i.e. Mobile IPv6 and SIP (Session Initiation Protocol). MIPv6 supports mobility at the network layer whereas Session Initiation Protocol (SIP) offers mobility at application layer and maintains session. Besides mobility, various problems of delay and packet loss need to be determined. To avoid these problems Fast MIPv6 [5] and Hierarchical MIPv6 [6] were introduced in which a user can be connected to more than one wireless networks at a time to acquire a smooth handover without disrupting the quality of ongoing session. In order to further minimize the handover delay, HeeYoung Jung et al. combined FMIPv6 and HMIPv6 to propose F-HMIPv6 [7]. Extensions to MIPv6 have been combined with SIP over different periods to improve QoS in NGNs [8].

The scope of this research is to provide a step forward in the provisioning of QoS and mobility transparency in NGNs. The proposed scheme of Integrating SIP with F-HMIPv6 (ISF) aims to handle the signal and traffic locally by creating a dynamic tunnel between mobility anchor point (MAP) and New Access Router (NAR), thus minimizing the extra burden on the core network.

The rest of the paper is organized as follow; section 2 deals with the background study and problem statement, section 3 discusses the proposed Integrated SIP with F-HMIPv6 (ISF), its algorithm and flowchart. Section 4 shows simulation results followed by section 5 describing conclusion and future work.

2 Related Work

2.1 Integration of SIP with MIPv6

Mobile IPv6 is considered as the Mobile IP network solution that allows the nodes to maintain reachability and ongoing connection as long as it remains connected to the Internet [8, 9]. MIPv6 avoids the concept of triangular routing and supporting the session continuity during handovers. However in this approach there is no discussion on handling multimedia session or any other delay sensitive applications [10].

2.2 Integration of SIP with SMIPv6

Integration of SIP with Seamless Mobile IPv6 (SMIPv6) was proposed to further improve QoS during handovers [3]. It reduced many problems like triangular routing, QoS while performing handovers but no detail and practical implementations are mentioned in the proposed scheme.

2.3 Integration of SIP with FMIPv6

Combining Session Initiation Protocol (SIP) with Fast Mobile IPV6 [11] in 4G networks provides real time mobility. It was proposed to reduce system redundancy and signaling exchange between edge and core networks. It handles the handover latency by foretelling and executing handovers in advance. This approach provides end-to-end Quality of Service (QoS) by using Advance Resource Management Techniques (ARMT). This technique proposes to use a QoS Manager for dynamic allocation of resources during handover when requested by the users. Problem in this approach is that every time binding updates need to go through core networks causing extra delay. Mobile Node (MN) has to wait for the acknowledgement which comes through the core network causing a half round trip delay before the packets are actually forwarded to the Correspondent Node (CN). The round-trip delay is greater when the MN is far away from the core network. Further, authentication between mobile node and correspondent node was handled by QoS manager which puts extra burden on QoS manager [10].

2.4 Integration of SIP with HMIPV6 (CSH)

The approach of Combining SIP with HMIPv6 (CSH) in NGNs provides terminal and session mobility, bandwidth management, resource reservation, user authentication and delay while establishing multimedia session [12]. However, the scheme only discusses latency due to binding updates without considering Movement Detection and CoA (Care of Address) configuration/ verification. Moreover extra delay occurs because traffic is routed through MAP (Mobility Anchor Point) [13].

The above discussion leads us to propose the following improvements regarding QoS in NGNs:

- Maintaining end-to-end QoS and reducing delay during multimedia session.
- Minimizing the signaling traffic and service disruption between edge and core networks.

3 Proposed Scheme

In this section, a new approach Integrating SIP with F-HMIPv6 (ISF) is proposed that merges F-HMIPv6--a network layer protocol with SIP-- an application layer protocol as shown in Fig. 1. The achievement of ISF ensures end-to-end QoS and minimizing delay, signaling traffic and service disruption. In order to consider fast handover a dynamic tunnel between MAP and MN is created in advance before the actual handover takes place.

The other module proposed is a QoS manager. The concept of QoS manager is taken from the research work proposed in integrating SIP with FMIPv6 [16] but with some improvements. QoS manager is responsible for managing QoS while users switching between different networks or between different regions under same network. QoS manager is focused on QoS parameters and not on the security concerns because security feature is provided by F-HMIPv6.

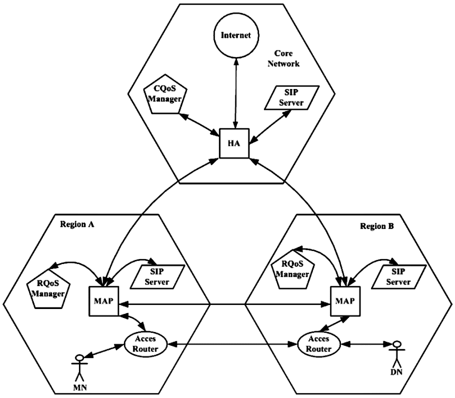


Fig. 1. Architecture of the proposed scheme

3.1 Session Initiation Protocol (SIP)

SIP is a text based and light weight application layer protocol. It provides signaling and control that is basically used for establishing, negotiating, modifying and terminating the ongoing sessions [14]. It is also used to support terminal mobility [15]. SIP is mainly concern with the session management and allows terminating of an existing or ongoing session. A user has to prior register with the SIP server before establishing a session. In the proposed scheme SIP is responsible to handle two kinds of Mobility:

Pre-Call Mobility. Pre-Call Mobility guarantees that during session or terminal mobility a CN can reach a MN. The MN, on changing its region sends its New Care of Address (NCoA) to the SIP Server of its home network. When the CN sends an INVITE message to the SIP server, it responds with the new CoA of MN. CN then sends an INVITE message to the MN and receives an Ok message from MN.

Mid-Call Mobility. Mid-Call Mobility assure the ongoing communication with its peer during handoffs. When MN initiates a handover during an ongoing session, it sends a Re-INVITE message with its New CoA. The CN replies with 200 Ok message in order to continue the session smoothly.

3.2 Fast Hierarchical Mobile Ipv6 (F-HMIPV6)

A new extension of MIPv6 called F-HMIPv6 was introduced to process the handovers in less interval of time. It is the combination of FMIPv6 and HMIPv6 that introduces the concept of creating a dynamic tunnel for fast handover between MAP and the New Access Router (NAR) [10]. MAP is a local Home Agent (HA) so there can be more than one MAP in a region. It provides the seamless connectivity by communicating with other networks before actual handover takes place. Furthermore it ensures the correct sequence of packet during handovers. It uses the concept of flush messages that helps in minimizing the delays. The main advantage of handling the mobility at network layer is that the applications are not aware of mobility.

3.3 QoS Manager

The QoS manager is responsible to allocate different resources on request of mobile users such as bandwidth allocation, implementing network policies and recourse reservation [3, 16]. In integration of SIP with FMIPv6 the security issues are handled by QoS manager [14], but in the proposed scheme it is handled by MAP which reduces the overhead on core network. In ISF architecture the QoS manager is classified on the basis of level of hierarchy as shown in Fig. 1.

Core QoS (CQoS) Manager. CQoS Manager is responsible to provide the requested resources to the Regional QoS (RQoS) manager, as their might be shortage of resources in their respected regions. CQoS manager has the ability to support handovers of different nature [14].

Regional QoS (RQoS) Manager. RQoS Manager is responsible for providing resources in their respective regions. They will maintain the communication with other RQoS managers when required during handovers.

3.4 Proposed Algorithm for Integrating SIP with F-HMIPV6 (ISF)

The proposed algorithm is designed for different networks and it includes two types of nodes: MN and Destination Node (DN). Other entity includes MAP, SIP server, Access Routers (ARs) and RQoS manager in every region. All the regions are connected with the core network. The core network has its own SIP server and CQoS manager.

ISF algorithm comprises of four different cases depending on four possible conditions that might occur during handover. The handover conditions include weak signal, user initiated handover or any other QoS issue. It also depends on session or network mobility or even both as per the user requirements.

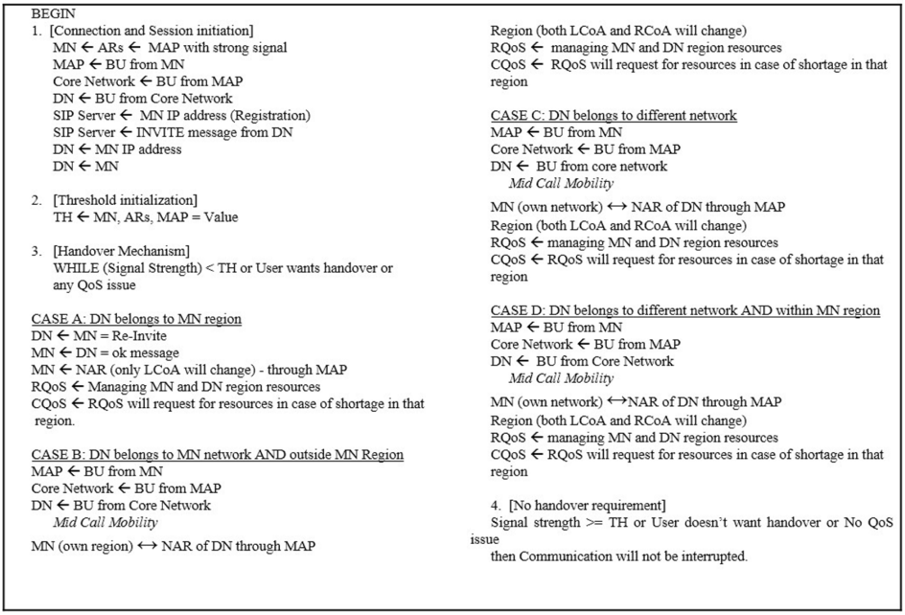
In first case, both the MN and the DN belongs to the same region and same network. If MN or DN changes their regions under same network only the on-Link Care of Address (LCoA) will be changed.

In second case, both the MN and DN belong to the same network but different regions. If the MN or DN changes its region, the handover takes place and both the LCoA and RCoA will change. The RCoA will be updated with both HA and DN.

In third case, both MN and DN belong to different networks and different regions. If MN or DN changes its region, both RCoA and LCoA will change during handover. RCoA will be updated with the HA of that particular network and also with DN.

In fourth case, both MN and DN belong to different networks but within same region. If MN or DN changes its point of attachment, both RCoA and LCoA will change during handover. RCoA will be updated with the HA of that particular network and also with DN.

The pseudo code for the proposed scheme of Integrating SIP with F-HMIPv6 (ISF) is given as under:



```
BEGIN
1.  [Connection and Session initiation]
    MN ← ARs ← MAP with strong signal
    MAP ← BU from MN
    Core Network ← BU from MAP
    DN ← BU from Core Network
    SIP Server ← MN IP address (Registration)
    SIP Server ← INVITE message from DN
    DN ← MN IP address
    DN ← MN

2.  [Threshold initialization]
    TH ← MN, ARs, MAP = Value

3.  [Handover Mechanism]
    WHILE (Signal Strength) < TH or User wants handover or
    any QoS issue

CASE A: DN belongs to MN region
DN ← MN = Re-Invite
MN ← DN = ok message
MN ← NAR (only LCoA will change) - through MAP
RQoS ← Managing MN and DN region resources
CQoS ← RQoS will request for resources in case of shortage in that
 region.

CASE B: DN belongs to MN network AND outside MN Region
MAP ← BU from MN
Core Network ← BU from MAP
DN ← BU from Core Network
    Mid Call Mobility
MN (own region) ↔ NAR of DN through MAP
Region (both LCoA and RCoA will change)
RQoS ← managing MN and DN region resources
CQoS ← RQoS will request for resources in case of shortage in that
region

CASE C: DN belongs to different network
MAP ← BU from MN
Core Network ← BU from MAP
DN ← BU from core network
    Mid Call Mobility
MN (own network) ↔ NAR of DN through MAP
Region (both LCoA and RCoA will change)
RQoS ← managing MN and DN region resources
CQoS ← RQoS will request for resources in case of shortage in that
region

CASE D: DN belongs to different network AND within MN region
MAP ← BU from MN
Core Network ← BU from MAP
DN ← BU from Core Network
    Mid Call Mobility
MN (own network) ↔NAR of DN through MAP
Region (both LCoA and RCoA will change)
RQoS ← managing MN and DN region resources
CQoS ← RQoS will request for resources in case of shortage in that
region

4.  [No handover requirement]
    Signal strength >= TH or User doesn't want handover or No QoS
issue
    then Communication will not be interrupted.
```

Complete architecture for the proposed algorithm is presented in Fig. 1. A MN in region-A wants to communicate with DN in region-B. MN will send the Binding Updates (BU) to MAP with strong signals in same region. There can be more than one MAP in a region; therefore MN will communicate with the MAP having strong signals. The BU includes both LCoA and RCoA of the MN. The connection of MN with MAP is handled through AR. The MAP will forward the BU to core network. The core network will then connect the MAP of region-A with the MAP of region-B for the first time. A session will be established between MN and DN through Core SIP server.

In case of pre call mobility, MN will send New CoA to SIP server of region-A. DN will send INVITE message to SIP Server of region-A. The region-A SIP server will forward the new CoA of the MN to the DN. The DN send INVITE message to the MN directly, and receives Ok message from MN. A session will be established known as pre call mobility.

After successful network and session connection, if handover occurs then the four different cases mentioned in the proposed algorithm will be considered to handle the network and session mobility. In any of the above discussed cases if the handover occurs, a tunnel will be created between MAP and NAR. The MAP will connect MN with DN and route the traffic between both the nodes through their ARs. A complete flow of integrating SIP with F-HMIPv6 (ISF) algorithm is presented in Fig. 2.

The RQoS manager is responsible to manage the resources in their respective region. If any shortage of resources occurs in any region a request will be made to CQoS manger. The CQoS manger will allocate resources in the requested region as per request by RQoS.

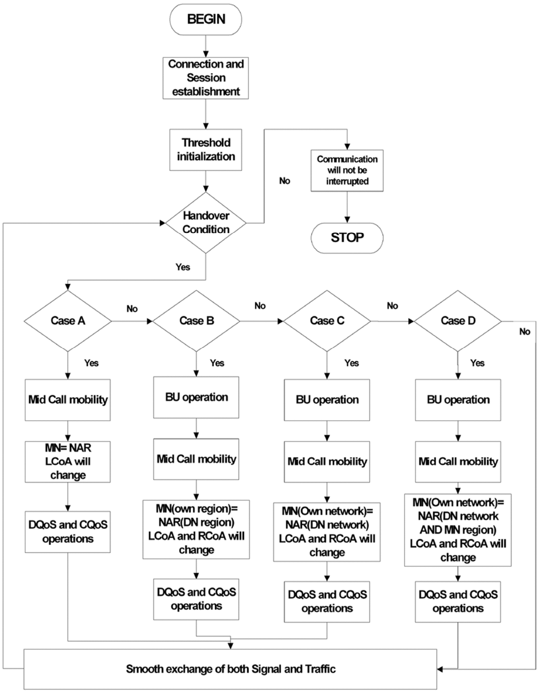


Fig. 2. Flow chart of the proposed scheme

4 Simulation Results

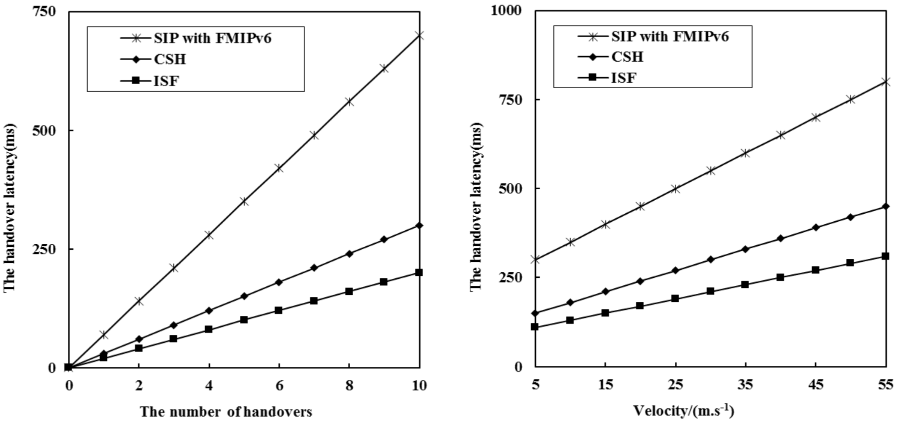
In this section, we compare our proposed scheme with the existing approaches such as integration of SIP with FMIPv6, and Combined SIP HMIPv6 (CSH). The simulation results in terms of minimizing latency, packet loss, and load are obtained through OPNET simulation tool. Various entities and nodes used in the simulation environment are presented in Fig. 1. The proposed scheme is evaluated in heterogeneous environment which includes diverse networks such as WiMAX, WLAN, and UMTS. Simulation parameters are demonstrated in Table 1.

4.1 Handover Latency

The handover latency obtained through simulation are presented is this sub section. Two cases are considered to evaluate the handover latency; one is handover latency versus velocity and the second one is handover latency versus number of handovers. In Fig. 3(a) number of handovers is kept variant for the evaluation of handover latency. Fig. 3(b) shows the increase in handover latency by increasing velocity. Our proposed scheme has 110ms to 310ms handover latency which is quietly reduced. The results demonstrates that our proposed scheme performs better than existing approaches.

Table 1. Simulation Parameters

Wired Bandwidth	1 Gb/s
Wireless link bandwidth	100 Mb/s
Packet size	1Kb
Moving speed	5-60 m/s
Radius of wireless cell	100 m
Simulation time	240 sec



(a) Handover latency versus number of handovers (b) Handover Latency versus Velocity

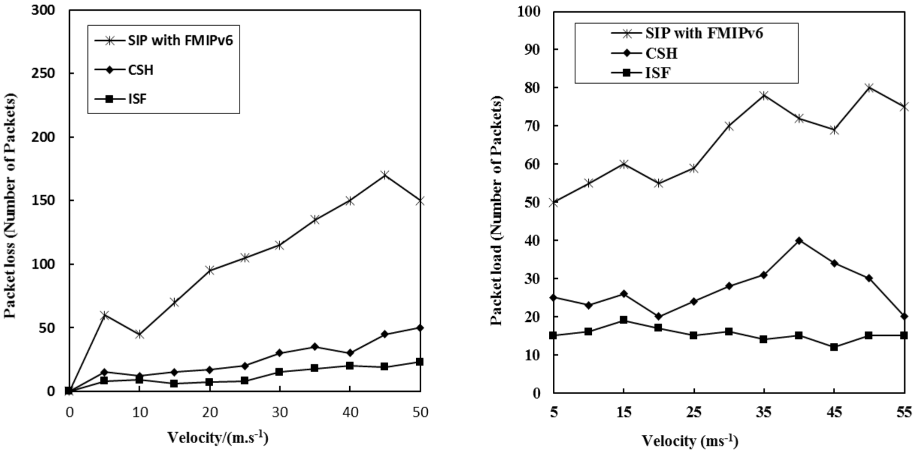
Fig. 3. Handover Latency

4.2 Packet Loss

During an ongoing session, we cannot ignore the loss of packets which can be occurred due to many reasons such as improper management of bandwidth, frequently handovers, and so on. In our simulations we evaluated packet loss regarding MN velocity. In Fig. 4(a), we can observe that our proposed scheme has lower packet loss as compared to the existing approaches. The maximum packet loss obtained for SIP with FMIPv6, CSH, and ISF are 150, 50, and 23 respectively. Hence, it is justified that ISF performance is best in terms of packet of loss.

4.3 Packet Load

The traffic and signal load is evaluated for our proposed scheme and existing approaches. The load based on packet is compared by increasing velocity. We can see from Fig.4 (b), ISF has lower packet load which is less than 20, whereas in CSH it is 40. Thus. In terms of packet load, our proposed scheme outperforms SIP with FMIPv6 and CSH.



(a) Packet Loss versus Velocity (b) Packet Load versus Velocity

Fig. 4. Packet loss/load versus velocity

5 Conclusions and Future Work

The researchers' focus is on many issues in NGNs such as wireless security, multicasting, resources allocation and management, QoS, and mobility transparency. Providing end-to-end QoS is still a challenge in heterogeneous environment. This paper proposed a new protocol Integrating SIP with F-HMIPv6 (ISF) which is a combination of an application layer protocol SIP and network layer protocol F-HMIPv6. For comparative analysis the ISF is compared with existing combination of SIP with FMIPv6, and SIP with HMIPv6. The analysis reveals that ISF aims to handle the signal and traffic locally by creating a dynamic tunnel between MAP and NAR thus minimizing the extra burden on the core network. The accomplishment of CSF includes end-to-end QoS, low signaling traffic between edge networks and core networks, to overcome the service disruption, user authentication, connection between MAP and MN is created in advance before the actual handover takes place.

In future the focus will be on security enhancement in the proposed architecture and efforts will be made to further enhance its efficiency.

References

1. Dong-Hoon, S., et al.: Distributed mobility management for efficient video delivery over all-IP mobile networks: Competing approaches. IEEE Network 27(2), 28–33 (2013)
2. Quintero, A., Frutos, E.D.: MPLS Based Architecture for Mobility and End-to-End QoS Support in Fourth Generation Mobile Networks. Journal of Computer Science 5, 255–262 (2009)
3. Faisal, S.: Performance Analysis of 4G networks. Department of Electrical Engineering School of Engineering Bleking Institute of Technology SE-37 79 Karlskrona, Sweden (2010)
4. Jesus, V., Sargento, S., Almeida, M., Corujo, D., Aguiar, R.L., Gozdecki, J., Carneiro, G., Banchs, A., Yáñez-Mingot, P.: Integration of mobility and QoS in 4g scenarios. In: Proceedings of the 3rd ACM Workshop on QoS and Security for Wireless and Mobile Networks (Q2SWinet 2007), pp. 47–54 (2007)
5. Johnson, D., Perkins, C., Arkko, J.: Mobility Support in IPv6. RFC3775, the Internet Society (June 2004)
6. Koodli, R. (ed.): Fast Handovers for Mobile IPv6. RFC4068, the Internet Society (July 2005)
7. Jung, H., Soliman, H., Koh, S.J., Takamiys, N.: Fast Handover for Hierarchical MIPv6 (F-HMIPv6). Internet Engineering Task Force (October 2005)
8. Li, Y., Zhao, Y.-S., Liu, Q.-L., Wen, F.: Performances Research of MIPv6 and extended Protocols in the process of Handover. In: Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom 2009), September 24-26 (2009)
9. Johnson, D., Perkins, C., Arkko, J.: Mobility Support in Ipv6 (MIPv6). RFC 3775, the Internet Society (June 2004)
10. Jung, H., Soliman, H., Koh, S.J., Lee, J.Y.: Fast Handover for Hierarchical MIPv6 (F-HMIPv6). Internet Engineering Task Force (April 2005)
11. Nursimloo, D.S., Kalebaila, G.K., Anthony Chan, H.: A two layered Mobility Architecture using Fast Mobile IPv6 and SIP. EURASIP Journal on Wireless Communications and Networking- Multimedia over Wireless Networks, 1–8 (2008)
12. Zubair, M., Mahfooz, S., Khan, A., Ur Rehman, W.: Providing end to end Qos in NGNs using combined SIP HMIPv6 (CSH). In: The Proceedings of the IEEE 1st International Conference on Computer Networks and Information Technology (ICCNIT 2011), July 11-13, pp. 113–118 (2011)
13. Soliman, H., Castelluccia, C., Elmalki, K., Bellier, L.: Hierarchical Mobile Ipv6 (HMIPv6) Mobility Management. RFC 5380 (October 2008)

14. Nursimloo, D.S., Chan, H.A.: Integrating fast mobile IPv6 and SIP in 4G network for real-time mobility. In: The Proceedings of the IEEE 13th International Conference on Networks, jointly held with the IEEE 7th Malaysia International Conference on Communication 2005, November 16-18, vol. 2, pp. 917–922 (2005)
15. NG, K.L.S.: Peer to Peer real time mobility using SIP and MIPv6. Electrical & Computer Engineering, National University of Singapore (2004)
16. Nakajima, N., Dutta, A., Schulzrinne, H.: Handoff Delay Analysis and Measurement for SIP based Mobility in IPv6. In: The Proceedings of the IEEE International Conference on Communications, ICC 2003, May 11-15, pp. 1085–1089 (2003)

Management of Inter-domain Quality of Service Using DiffServ Model in Intra-domain

Sara Bakkali, Hafssa Benaboud, and Mouad Ben Mamoun

LRI, Faculty of Sciences at Rabat, Mohammed V-Agdal University,
Rabat, Morocco
bakkalisara@gmail.com,
{benaboud,ben_mamoun}@fsr.ac.ma

Abstract. During the last decade, Internet has experienced enormous evolution. This evolution concerns the huge quantity of traffic circulating over Internet and also the important diversity of these traffics types. Each type of traffic requires a specific QoS parameters. This point may represent a serious concern mainly due to the difficulty in ensuring QoS for traffics that cross multiple domains or Autonomous Systems (ASs). To solve this problem several researchs and studies has been proposed. In this paper, we describe a new mechanism that we proposed in a previous work to solve this problem. Our mechanism ensures the end to end QoS requirements over multiple ASs. This method keeps the same values of QoS parameters required by the traffic, even during its passage across several ASs. This paper explains the problem of inter-domain QoS, describes our new approach, we give then a case study of our solution using DiffServ model in intra-Domain. Simulation Results show the improvement of network performance when using the new mechansim and when comparing them to those of the standard case.

Keywords: Inter-domain routing, QoS, DiffServ.

1 Introduction

Today, traffics circulating on networks are very diversified and require a specific parameters in terms of bandwidth, delay and other necessary parameters. View the limited network resources, it was necessary to find a mechanism for a QoS management within a network. To solve this problem a various models have been implemented to ensure QoS whithin the same network or in intra-domain case. However, in Internet which is an inter-domain network the problem is not resolved yet. In this context we present this paper which describes in details the new method that we prorposed in [1]. This method ensures the end-to-end QoS constraints for services across multiple domains. Services involved in our approach include real time services such as voice and video telephony and conference, as well as services those requiring high capacity interconnections like links between scientific sites or cloud services, which are provided by different domains or AS's.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_69, © Springer International Publishing Switzerland 2014

The remainder of this paper is organized as follows. In section 2 we present related works and define the inter-domain QoS problem. Next in section 3, we describe our approach that ensures end-to-end QoS over multiple domains. Then in section 4, we present a simulation of our approach using DiffServ model. And finally, in section 5, we conclude this paper and give future works.

2 Related Works and Inter-domain QoS Problem

2.1 Inter-domain Problem

Several solutions and technologies have been proposed and implemented to provide QoS within the same domain (AS), such as IntServ (Integrated Services) [2] model, DiffServ (Differentiated Services) [3] model or even MPLS [4]. However, a serious problem is posed when the traffic crosses another domain (AS). This problem is mainly due to the fact that QoS constraints, required by the client and which the operator undertakes to provide (usually specified in the Service Level Agreement, SLA), are defined in the classes of service. While the definition of the classes of service is assured by the domain administrator, they are consequently specific to each domain, and are valid only within this domain. In this case, in the transition to another domain the QoS constraints offered to the traffic will not be the same as in the source domain, therefore the QoS required by the client at the beginning will not be provided from the end-to-end until its destination.

2.2 Related Works

A various studies and several solutions have been proposed to ensure QoS in inter-domain; each solution suggests a specific approach to treat the problem. Among these solutions, is an extension of traffic engineering in MPLS (Multi-Protocol Label Switching) architecture for inter-domain's use called inter-domain MPLS Traffic Engineering [5]. This solution is mainly based on the label's exchange between edge routers, and bandwidth reservations using enhanced version of Resource Reservation Protocol (RSVP). Also, MESCAL project (Management of End-to-end QoS across the Internet At Large) [6] introduces a new architecture for inter-domain QoS management. However, it focuses only on financial management between customers and operators and between operators. Likewise, a complete model has been proposed in [7] to provide management functionality for End-to-End QoS by combining a routing procedure, a common set of QoS operations and an information model. However, it's specific to dedicated point-to-point connections. Authors of [8] and [9]treated the path computation aspect of the inter-domain QoS routing by providing a new algorithm named HID-MCP (Hybrid Inter-Domain Multi-Constraint Path for inter-domain multi-constraint QoS paths computation. Nevertheless, this solution concerned only paths pre-computation or computation, and didn't offer a complete approach to ensure end-to-end inter-domain QoS.

All these inter-domain solutions do not provide to client's traffic the same QoS required as in its source domain. In this context, we introduce this paper which presents a solution that offer to client's traffic the same QoS constraints even in passing to another domain.

3 Proposed Solution Description

3.1 Approach Definition

To solve the problem mentioned above, and to ensure continuity of QoS constraints offered to the client even after the transition to other domains, we introduce a new method that provides a new mechanism for inter-domain traffic treatment. The basic idea in our approach is to designate in each domain a server responsible for the management of the different classes of service, named the **Class Manager** (CM). On this server we define a table, named **Class Table** (CT) that contains all information concerning the different classes defined in this domain (such as bandwidth, loss rate, delay, etc.). Once the CM of each domain filled its CT, it sends it to the neighbouring domain. In this way, each CM has all the information about its neighbours classes of services, and then, upon receiving a packet from the neighbouring domain, the router in the current domain can classify it in a class that has the same characteristics as the source class. In this manner, the client flow retains the same QoS constraints throughout its path to the destination, and receives the same treatment from end-to-end.

3.2 CT Table Structure

The class table is structured according the following fields:

1. AS number: to identify domain associated with the class.
2. Class number: to identify the class of service.
3. Bandwidth: to indicate the percentage of bandwidth allocated to the class.
4. Priority: to specify the priority level of the class.
5. Queue-limit: to specify the maximum number of packets that the queue can hold for this class.

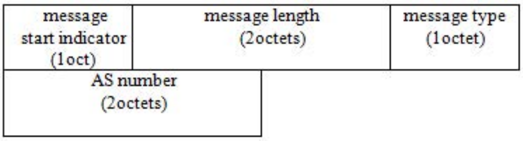
We note that, to ensure a certain correspondence between the CT tables of the different domains, we define in the CT table only class parameters common between various router's constructors, which are basic parameters used by the different constructors to characterize a class of service, other parameters more specific and appropriate for each router's constructor are not considered in the CT table. The parameters used in the CT table must be specified in the agreement established between the domains as we will explain later in this paper.

3.3 Sending Information from Routers to CM Sever

As we have already mentioned, routers classify the customer traffic by applying mechanisms of the adopted QoS intra-domain model. Informations concerning parameters relatives to every class defined on a router are in the router configuration file. We propose that the routers execute a PerlScript to retrieve information concerning classes of service from the configuration file, and to place them in a new file. This file will be sent to the CM server. Once the CM server receives all routers files, it regroups them in a file named CT, that represent the class table in which are stored informations concerning all classes of service defined in the domain.

3.4 Exchanging Tables between CM Servers

The communication between all domains CM servers uses the TCP protocol. So, in order that a CM sever cen send its CT table to the neighbouring domain CM server and receive its CT, they establish at first a TCP session. Once the session TCP is established, the first message exchanged between both CM servers is the identification message, which allows each CM server to become identified by its neighbour, by sending its IP address and AS number. The identification message format is presented in the following figure:



message start indicator (1oct)	message length (2octets)	message type (1octet)
AS number (2octets)		

Fig. 1. Identification Message Format

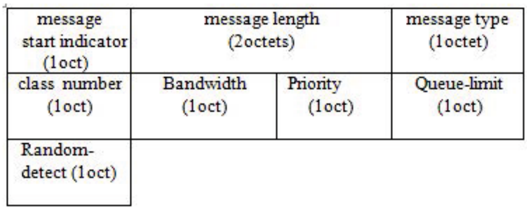
After the identification, CM servers exchange their CT tables by sending a set of messages to announce their classes of services, called announcement messages. Every message contains various parameters values relatives to every class defined in the domain.

The format of every message is as follows:

Information contained in every message as soon as it's received by the CM server it's registered in its CT table. This way when the CM server receives the totality of messages, it will have all information concerning all classes defined in the neighbouring domain.

The last type of message is the update message, which is sent by a CM server when there is an addition or modification of a class of service defined in its domain. The update message has the same structure as the announcement message.

Once a CM server receives its neighbour CT table, it diffuses it to the routers of its domain. Hence, all domain routers will possess all information about classes



message start indicator (1oct)	message length (2octets)		message type (1octet)
class number (1oct)	Bandwidth (1oct)	Priority (1oct)	Queue-limit (1oct)
Random-detect (1oct)			

Fig. 2. Announcement Message Format

of service defined in the neighbouring domain, and can use this information to create and configure classes of service which will have same values of QoS parameters. According to these classes of service packets coming from the neighboring domain will be classified with the same QoS constraints and will be forwarded in the current domain.

3.5 Agreements between Domains

The proposed solution is mainly based on agreements established between domains. Indeed, the information exchanged between domains in CT tables is very important and very sensitive information and the domain administrators have to negotiate and establish an agreement that will manage relations between domains so that the exchange of CT tables takes place with no problem. The agreement also defines how the table's exchange will be charged.

4 Simulation Using DiffServ Model in Intra Domain

After the description of our approach in the previous sections, we present a simulation of a sample application to better understand the approach and its operation, and also to prove the efficiency of its principle. In this example we treat the case where the network uses DiffServ model to provide QoS in intra-domain and we consider a client with sensitive, important and expensive application that needs to use the resources in the neighbouring domain. Firstly, we will briefly present DiffServ model, to describe then the architecture and the scenario of the simulation.

4.1 DiffServ Model

DiffServ is a service model that ensures the QoS requirements in a network. The client's flows in a network are treated by creating differentiated service classes with different priorities[10]. The main advantage of DiffServ over other models, is its simplicity and robustness especially in large-scale implementations. This robustness is due to the fact that in DiffServ are two types of routers: routers in the network core (core router) and the edge routers, only the edge routers that handle complex treatment of the flows that require resources and consume bandwidth and time. The edge routers perform classification, control and marking operations, and calculate an 8-bit DSCP (DiffServ Code Point) label that indicates the packet's class of service[11].

As mentioned above, we consider a client with a sensitive, important and expensive application that needs to use the resources in the neighboring domain. The use of DiffServ model allows classifying the client traffic in a class of service that responds to all the required QoS constraints, but only within its AS source. However, in some cases, this sensitive traffic must pass to the neighbouring domain and then, it loses the QoS values assigned to it in its own domain as agreed. The Requiered QoS is not provided from end-to-end. Obviously, in DiffServ model, the definition of classes of service is assured by the domain administrator, so, they are specific to each domain, and are valid only within this domain.

4.2 Simulation Topology Description

To solve the problem mentioned above we use our approach proposed in Section 3. Then, this simulation objective is to show performances of this approach, which consists in the fact that the user traffic is classified in classes of service with the same parameters even if they are located in two different domains. For that, we use the network simulator ns2 to compare two simulation scenarios, in both cases we consider two networks that use the diffserv model for QoS management in intra-domain, in the first scenario the two networks use classes of service with different parameters (it is the case in conventional networks), and in the second case both networks use same parameters for their classes of service (which is the principle of our new method). The topology we simulate is presented in figure 3.

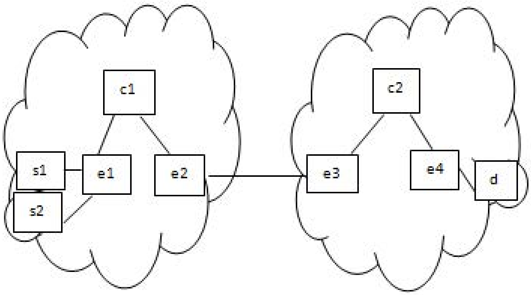


Fig. 3. Simulation Topology

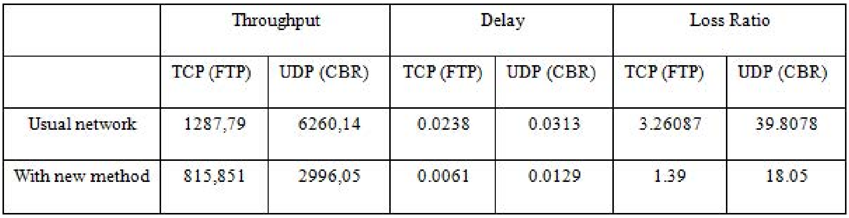
Simulation parameters in the first case are the following: On the node s1 tcp agent is configured to emit ftp traffic, and on the node s2 an udp agent is configured to send cbr traffic. In the first network we define two classes of service, the first one with the DSCP code 10, in which we classify the tcp traffic and the second with the code DSCP 11 in which we classify the udp traffic. The queue size of the two classes is 50 packets, they have two levels of priority (virtual queue), and a token bucket policer with CIR=100 kbps(Committed Information Rate) and CBS=10bytes (Committed Burst Size) for the first class, and CIR=300 kbps and CBS=40 Kbytes for the second class. In the second network we also define two classes of service but with different parameters, the first with DSCP 10 in which we classify the tcp traffic and the second with DSCP 11 in which we classify the udp traffic. The queue size of both classes is 20 packets, they have a two levels of priority (virtual queue), and a token bucket policer with CIR=1 Mbps (Committed Information Rate) and CBS=3 Kbytes (Committed Burst Size) for the first class, and a CIR=3 Mbps and CBS=10 Kbytes for the second class.

Simulation parameters in the second case are the following: On the node s1 tcp agent is configured to emit ftp traffic, and on the node s2 an udp agent is configured to send cbr traffic. In each network we define two classes of service; the two classes defined in the first network have the same parameters as those defined in the second one (to respect the principle of our method). The first class is defined with DSCP code 10 in which the tcp traffic is classified and the second class is defined with DSCP code 11 in which the udp traffic is classified. The queue size of both classes is 50, they both have two levels of priority (virtual queue), and a token bucket policer with CIR=100 kbps and CBS=10 bytes for the first class, and CIR=300 kbps and CBS=40 Kbytes for the second class.

4.3 Simulation Results

By simulating the architecture already presented with the parameters that we have detailed above, we obtain the results listed in figure 4. These results concern the end-to-end calculation of three main parameters to estimate the network performances; the throughput, the delay and the loss rate. By analyzing the results presented in the figure 5, which represent the average values of the different calculated parameters, we note that the use of the new method principle (the second simulation case) decreased significantly the end to end throughput, which means a decrease of the end to end link utilization rate for both types of traffic (tcp and udp) that allows a better optimization of network resources while improving the conditions for routing traffic since delay and loss rate also decreased. We also plot the instantaneous variation of the previous parameters, to compare the two scenarios.

The figures 5 and 6 represent the throughput variation in function of time. During all the duration of simulation (6 seconds), we note that the throughput values in the second case of simulation (which represents the new method) are lower than those of the first case (which represents an ordinary network). According to the figures 7 and 8 we also note a significant decrease in the instantaneous loss rates values for both types of traffic (tcp and udp) comparing the second case simulation with the first case.



	Throughput		Delay		Loss Ratio	
	TCP (FTP)	UDP (CBR)	TCP (FTP)	UDP (CBR)	TCP (FTP)	UDP (CBR)
Usual network	1287,79	6260,14	0.0238	0.0313	3.26087	39.8078
With new method	815,851	2996,05	0.0061	0.0129	1.39	18.05

Fig. 4. End to End Average QoS Values

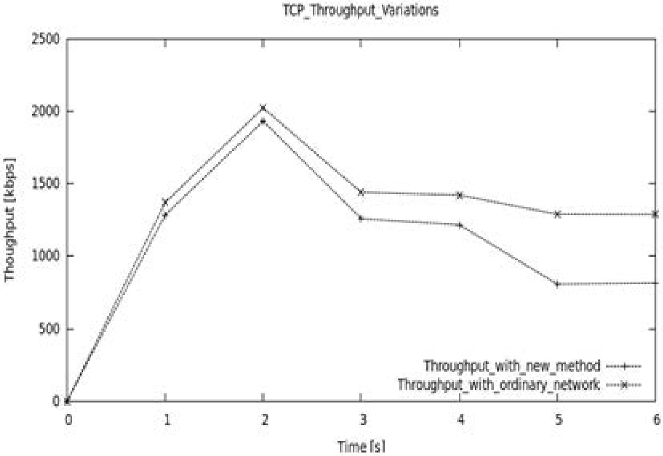


Fig. 5. TCP Instantaneous Throughput Variations

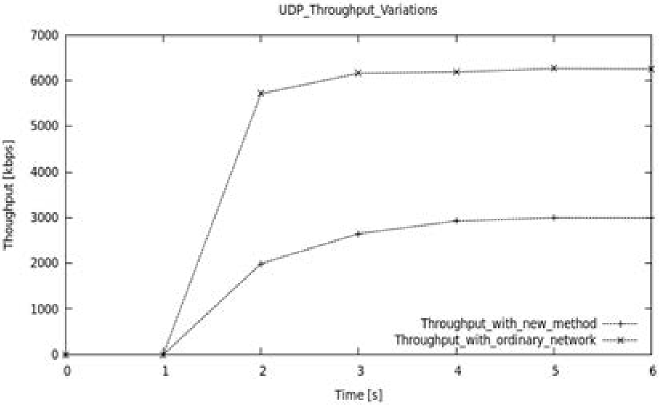


Fig. 6. UDP Instantaneous Throughput Variations

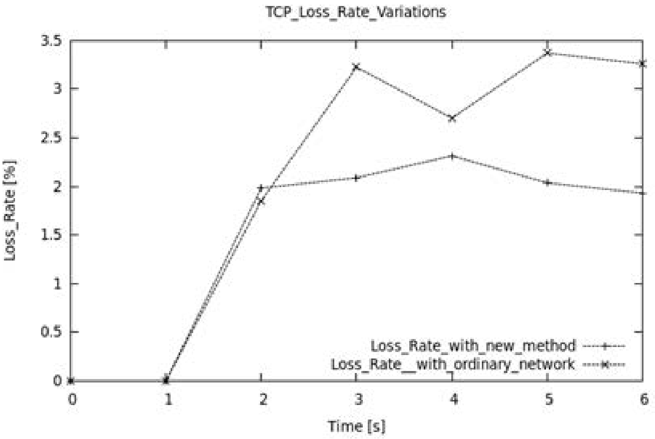


Fig. 7. TCP Instantaneous Loss Ratio Variations

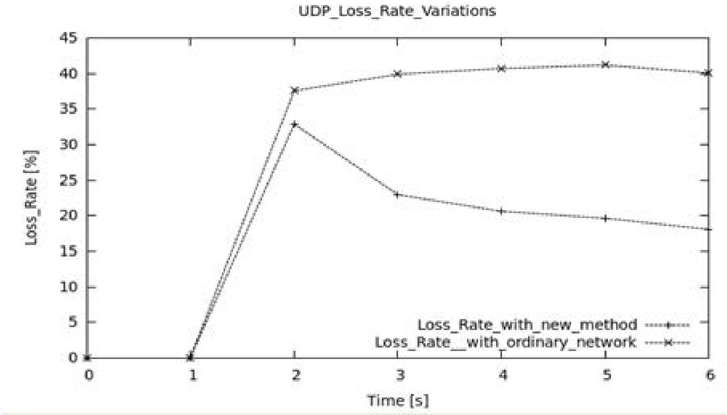


Fig. 8. UDP Instantaneous Loss Ratio Variations

We note that the simulation results are illustrative, and allow us to deduce that the use of our new method principle; which consists on keeping the same QoS parameters even in another domain; has improved network performance by reducing the delay and loss rate and also has ensure a better optimization of network resources by reducing the utilization rate of the end to end link.

5 Conclusion and Future Work

In this paper, we proposed a new mechanism which ensures end-to-end QoS over multiple AS. We described it and we gave details of its operation and its components. We gived then a simulation of this approach using the DiffServ model for providing QoS in intra-domain network. Simulation Results show the improvement of network performance when using our mechanism and of course, traffic keeps the same QoS provided by its own domain when it is destined to another AS. The next step of our research will focus on evaluating performance of the proposed approach in other environments by taking into account various models proposed in intra domain, and also on studying and proposing a mechanism for securing this approach.

References

1. Bakkali, S., Benaboud, H., Ben Mamoun, M.: On Ensuring End-to-End Quality of Service in Inter-Domain Environment. In: Gramoli, V. (ed.) NETYS 2013. LNCS, vol. 7853, pp. 326–330. Springer, Heidelberg (2013)
2. Shenker, S., Partridge, C., Guerin, R.: Specification of Guaranteed Quality of Service. IETF Informational, RFC 2212 (September 1997)
3. Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., Weiss, W.: An Architecture for Differentiated Services. IETF Informational, RFC 2475 (1998)
4. Rosen, E., Viswanathan, A., Callon, R.: Multiprotocol Label Switching Architecture. IETF Informational, RFC 3031 (2001)
5. Farrel, A., Vasseur, J.-P., Ayyangar, A.: A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering. IETF Informational, RFC 4726 (2006)
6. Howartha, P., Boucadairb, M., Flegkasa, P., Wanga, N., Pavloua, G., Morandb, P., Coadicb, T., Griffinc, D., Asgarid, A., Georgatsosen, P.: End-to-end quality of service provisioning through inter-provider traffic engineering. Computer Communications 29, 683–702 (2006)
7. Yampolskiy, M., Hommel, W., Danciu, A., Metzker, G., Hamm, K.: Management-aware Inter-Domain Routing for End-to-End Quality of Service. International Journal on Advances in Internet Technology 4, 60–77 (2011)
8. Frikha, A., Lahoud, S., Cousin, B.: Hybrid Inter-Domain QoS Routing with Crankback Mechanisms. In: Balandin, S., Koucheryavy, Y., Hu, H. (eds.) NEW2AN 2011/ruSMART 2011. LNCS, vol. 6869, pp. 450–462. Springer, Heidelberg (2011)
9. Frikha, A., Lahoud, S.: Hybrid Inter-Domain QoS Routing based on Look-Ahead Information. IRISA's Interne Publications, PI 1946 (2010)
10. Serban, R.: La gestion dynamique de la qualit de service dans linternet, archives inria el-00408686, version 1, Universit de NICE SOPHIA-ANTIPOLIS UFR SCIENCES (2003)
11. Nichols, K., Blake, S., Baker, F., Black, D.: Definition of the Differentiated Services Field (DS Field)in the IPv4 and IPv6 Headers. IETF Standars, RFC 2474 (1998)

Bilinear Representation of Non-stationary Autoregressive Time Series

Ewa Bielinska

The Silesian Technical University, Gliwice, Poland

Abstract. This paper considers a class of non-stationary autoregressive systems in which non-stationarity is caused by time varying parameters of the system. Distinction between two or more non-stationary systems based on observation of the output signal only, is difficult and sometimes may be impossible. In this paper a bilinear approximation of non-stationary autoregressive model is proposed. This way, a model with time varying parameters is approximated by a constant parameters model, what can facilitate the distinction between systems.

Keywords: non-stationary AR models, bilinear approximation, identification, system recognition.

1 Introduction

Modeling non-stationary time series has been a difficult task in spite of parametric or nonparametric methods. In time series analysis, autoregressive models are widely applied amongst researchers because they define linear regression, thats why model parameters can be easily estimated. Originally, AR models are intended for modeling stationary dynamic systems, however they are also used in modeling non-stationary systems. Generally, idea of modeling of the non-stationary process is to consider the process to be locally stationary, and fit stationary AR models locally. In such a way, a global non-stationary time series is represented by a dynamic set of local linear regressive models, named a treshold model. Several attitudes to this problem have been published, e.g. [9], [8], [6]. When the local AR models have the same structure, the treshold model may be interpreted as AR model with time varying parameters.

Presented research aims at distinction signals coming from different non-stationary systems. For example, this is a case of speaker recognition on the base of a registered utterance. The utterance itself is a time series that is modeled as non-stationary auto regressive (AR) series. In speech processing, analyzed utterance is segmented into frames in which stationary AR(dA) is identified. Then the models' parameters are processed to obtain the speech or the speaker features.

In the paper, a different attitude is proposed. Non-stationary AR model is approximated by a constant parameters bilinear model. In the subsequent sections derivation of the bilinear approximation, a method of bilinear parameters identification and possible applications are discussed.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_70, © Springer International Publishing Switzerland 2014

2 Bilinear Representation of Non-stationary Autoregressive Time Series

Consider a non stationary time series $s(i)$, for $i = 1 \dots N$ with the auto-regressive $AR(dA)$ representation

$$s(i) = \sum_{j=1}^{dA} a_j(i)s(i-j) + \nu(i) \tag{1}$$

where $a_j(i)$ for $j = 1 \dots dA$ are time varying parameters of the $AR(dA)$ model, and $\nu(i)$ is a series of errors, distributed $N(0, \sigma^2)$.
Assume that parameters $a_j(i)$ change in the following way

$$a_j(i) = a_j(i-1) + \Delta_j(i) \tag{2}$$

where Δ_j is a random innovation signal, with the moving average $MA(dC)$ representation

$$\Delta_j(i) = e_j(i) + \sum_{l=1}^{dC} c_{l,j} e_j(i-l) \tag{3}$$

where $e_j(i)$, $j = 1 \dots dA$ are identically independently distributed normal variables, with mean 0 and variance λ_j^2, and $c_{l,j}$ are constant, real parameters for $j = 1 \dots dA$ and $l = 1 \dots dC$
Under the assumption that initial value at time 0 is $a_j(0) = \Delta_j(0)$, equation (2) is equivalent to

$$a_j(i) = \sum_{k=0}^{i} \Delta_j(k) \tag{4}$$

Under (3), $a_j(i)$ can be rewritten as

$$\begin{aligned} a_j(i) = \sum_{k=0}^{i} e_j(k) + \sum_{k=0}^{i} \sum_{l=1}^{dc} c_{l,j} e_j(k-l) = \\ \sum_{k-0}^{i} e_j(k) + c_{1,j} \sum_{k=0}^{i} e_j(k-1) + c_{2,j} \sum_{k=0}^{i} e_j(k-2) + ... + c_{dC,j} \sum_{k=0}^{i} e_j(k-dC) \end{aligned} \tag{5}$$

Therefore, non-stationary autoregressive representation (1) can be rewritten as

$$\begin{aligned} s(i) = \sum_{j=1}^{dA} s(i-j) \left(\sum_{k=0}^{i} e_j(k) + \sum_{k=0}^{i} \sum_{l=1}^{dc} c_{l,j} e_j(k-l) \right) + \nu(i) = \\ \sum_{j=1}^{dA} \big(s(i-j)e_j^*(i) + c_{1,j} e_j^*(i-1)s(i-j) + ... + c_{dC,j} e_j^*(i-dC)s(i-j) \big) + \nu(i) \end{aligned} \tag{6}$$

where

$$e_j^*(i) = \sum_{k=0}^{i} e_j(k)$$

$$e_j^*(i-l) = \sum_{k=0}^{i} e_j(k-l)$$

From the above one can see that the non-stationary signal $s(i)$ represented previously by dA time varying coefficients $a_j(i)$, under assumption (2) can be also represented by $(dA \times dC)$ time constant coefficients $c_{j,k}$ of the bilinear model (6). The set of the bilinear model's parameters is gathered in the matrix $\mathbf{C}((dA \times (dC+1))$.

$$\mathbf{C} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,dC} & \lambda_1^2 \\ c_{2,1} & \cdots & c_{2,dC} & \lambda_2^2 \\ \vdots & \vdots & \vdots & \vdots \\ c_{dA,1} & \cdots & c_{dA,dC} & \lambda_{dA}^2 \end{bmatrix} \tag{7}$$

Bilinear models are linear with respect to the model's parameters what enables us to estimate the unknown parameters $c_{j,k}$ with the use of LMS or GLMS identification method. However, in this particular case, the regression vector is build of the products of $s(i-j)$ and $e_j^*(i-k)$, for $j = 1 \ldots dA$ and $k = 0 \ldots dC$. The $s(i-j)$ is known, but $e_j^*(i-k)$ can be neither measured nor easily estimated from the possessed data, and problem with identification of the model (6) arises.

3 A Tricky Idea of the Model Parameters Estimation

The bilinear representation of the non-stationary $AR(dA)$ time series (6) can be presented as

$$s(i) = \phi(i)\Theta + \nu(i) \tag{8}$$

where $\phi(i)$ is the regression vector

$$\begin{aligned} \phi(i) = [&f_0;\ s(i-1)e_1^*(i-1);\ s(i-1)e_1^*(i-2); \ldots ; s(i-1)e_1^*(i-dC); \\ &s(i-2)e_2^*(i-1);\ s(i-2)e_2^*(i-2); \ldots ; s(i-2)e_2^*(i-dC);\ \ldots \\ &s(i-dA)e_{dA}^*(i-1);\ s(i-dA)e_{dA}^*(i-2);\ \ldots ; s(i-dA)e_{dA}^*(i-dC)\ ] \end{aligned} \tag{9}$$

The first term f_0 in the regression vector is

$$f_0 = \sum_{j=1}^{dA} s(i-j)e_j^*(i) \tag{10}$$

Θ is vector of unknown parameters

$$\Theta = [1\ c_{1,1}\ c_{2,1}\ \dots c_{dC,1}\ c_{1,2}\ c_{2,2}\ \dots c_{dC,2}\ \dots c_{1,dA}\ c_{2,dA}\ \dots c_{dC,dA}]^T \tag{11}$$

Taking into account, that the regression vector (9) contains information that is not available in practice, we have to propose an alternative, indirect algorithm of estimation of model's parameters $c_{j,k}$. The algorithm requires the values of the $s(i)$ only.

1. Assume an order of non-stationary $AR(dA)$ model of the non-stationary time series $s(i)$
2. Estimate the time varying coefficients $a_j(i)$ for $j = 1 \dots dA$ of the linear model of $AR(dA)$ using e.g. recursive RLS method
3. Calculate the innovations $\Delta_j(i)$ as $\Delta_j(i) = a_j(i) - a_j(i-1)$
4. Calculate autocorrelation function $R_{\Delta_j(\tau)}$ for each of the innovation series $\Delta_j(i)$
5. Estimate $c_{j,k}$ coefficients using e.g. autocorrelation function $R_{\Delta_j(\tau)}$. Idea of MA time series estimation and two numerical procedures were presented in in [2]. Since then many algorithms for MA parameter estimation have been developed e.g. [5], [7], [4], [11].

4 Possible Applications of the Proposed Models

In this section, we will consider possible application of the proposed bilinear model (6) of the non-stationary time series. In general, the model is nonlinear,but linear with respect to the constant parameters. Signals $s(i-j)$ $j = 1 \dots dA$, and $e_j^*(i-k)$ for $k = 0 \dots dC$, are the model's inputs. Bilinear model is an alternative description of the linear time series with time varying parameters. For the sake of inaccessibility of the model's inputs $e_j^*(i)$, the model can hardly be used for prediction or control. However, it can be used to classify and recognize signals coming from different systems, because it is easier to conclude comparing respective constant parameters (c_j) than comparing respective time series $(a_j(i))$. The (c_j) parameters can be interpreted as information about dynamic character of $(a_j(i))$ course.

4.1 Discrimination of Signals Coming from Two Non-stationary Systems

Consider two non-stationary autoregressive $AR(2)$ systems, with time varying autoregressive parameters. An example course of time series coming from one of them is illustrated in the Fig.1.

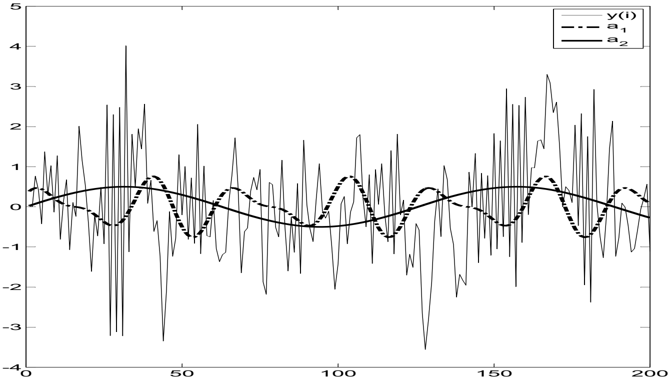


Fig. 1. Non-stationarity of AR(2) time series y(i) results from parameters a_1 and a_2 varying in time

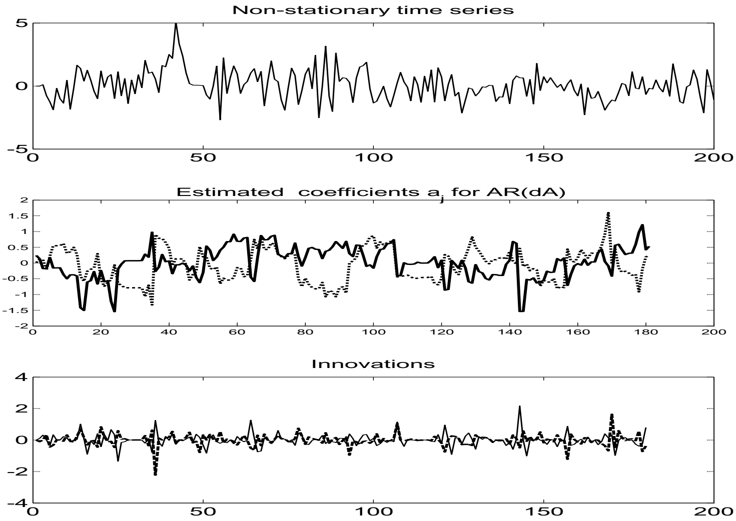


Fig. 2. Estimated time varying autoregressive parameters $a_1(i)$ and $a_2(i)$ of non-stationary time series $y(i)$, and their innovations

We want to investigate if the proposed method let us to know which of the non-stationary systems generated the observed time series y(i). To this aim, firstly we estimate the time varying coefficients $a_j(i)$ for $j = 1 \ldots dA$ of the linear model of $AR(dA)$ using recursive RLS method. An example result obtained for a time series coming from the second system is presented in the Fig.2.

At the top we can see non-stationary time series course, in the middle there are estimated time varying parameters $a_1(i)$ and $a_2(i)$, and at the bottom we have innovations $\Delta_1(i)$ and $\Delta_2(i)$.

In the last step we estimate $MA(2)$ models for innovation $\Delta_1(i)$ and $\Delta_2(i)$. The moving average parameters are, at the same time, the parameters of the bilinear model (6). In the Fig.3 there are shown estimated parameters c_1 and c_2 for the courses of $a_1(i)$ – at the top; and $a_2(i)$ – at the bottom.

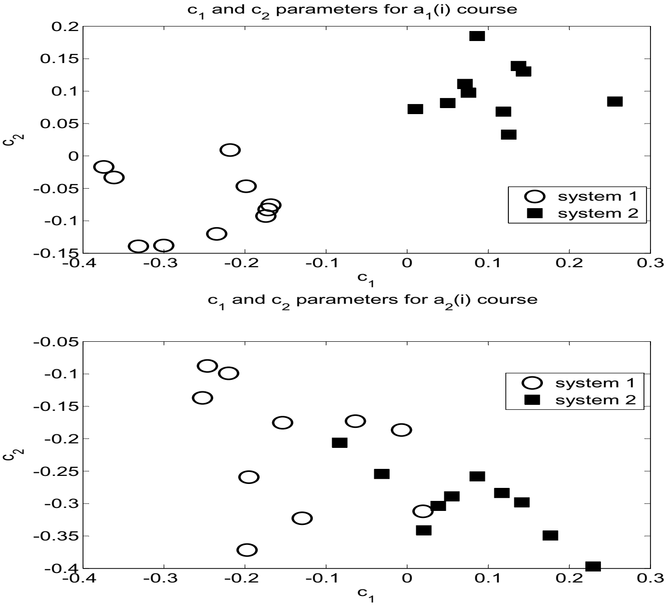


Fig. 3. Estimated parameters of bilinear models for two different non-stationary systems

Squares indicate results for one of the system, and circles – for the other.

We can see that the sets are separable and therefore, we will be probably able to recognize the system which has generated the observed time series. The next figure let us observe effect of identification of four stationary AR(2) systems with the same time invariant parameters but with different stochastic term. We can see that the diagrams overlay, what indicates that the sets are not separable, and the analyzed data come from the same dynamic system.

4.2 Speaker Recognition

An utterance generated by a speaker can be characterized by a set of features that describe as well the speaker as his utterance. With the aim of speaker recognizing,

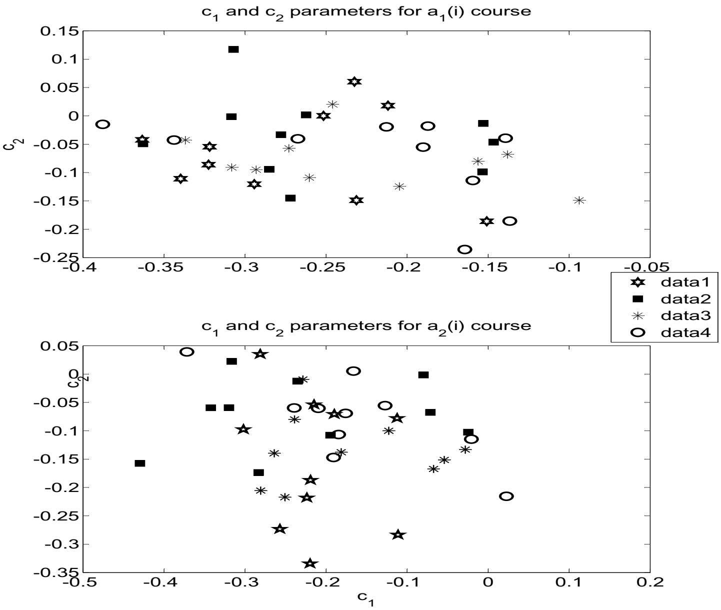


Fig. 4. Estimated parameters of bilinear models for four stationary systems with the same autoregressive parameters and different stochastic term

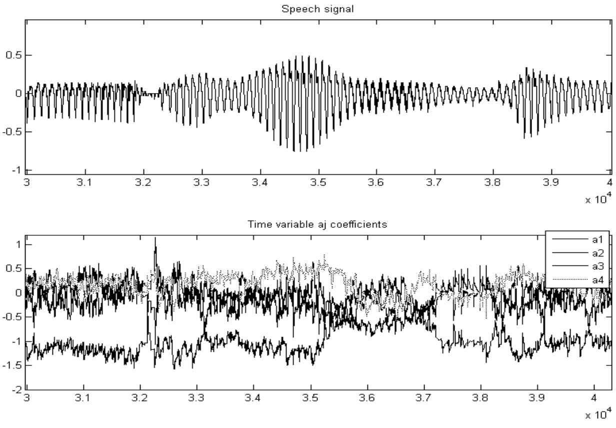


Fig. 5. Speech signal and estimated courses of four autoregressive parameters

the features which characterize the speaker only, and not the specific fragment of text should be distinguished from the registered fragment of utterance. The basic methods of speaker recognizing (e.g. [1], [3]) consist in comparison of the speaker features averaged over frame boxes, on which the registered signal is divided, during a test session, to a pattern in a speaker base. Under usual taken

assumption that speech signal is linear non-stationary, we can find its bilinear representation. The feature vector is define as set of constant bilinear coefficients c_j. Number of coefficients depends on assumed model structure.
In this research, silence was first removed from the registered utterance, and such modified signal was described by non-stationary AR(4) time series model (1). Then, the time varying coefficients $a_j(i)$ of the AR(4) were estimated with the use of RLS. Fig.5 shows example fragment of utterance, and estimated courses of four autoregressive parameters

Next, according to the presented above algorithm, innovation $\Delta_j(i) = a_j(i) - a_j(i-1)$ and their autocorrelation functions $R_{\Delta_j}(k)$ were calculated. Example results are illustrated in the Fig.6.

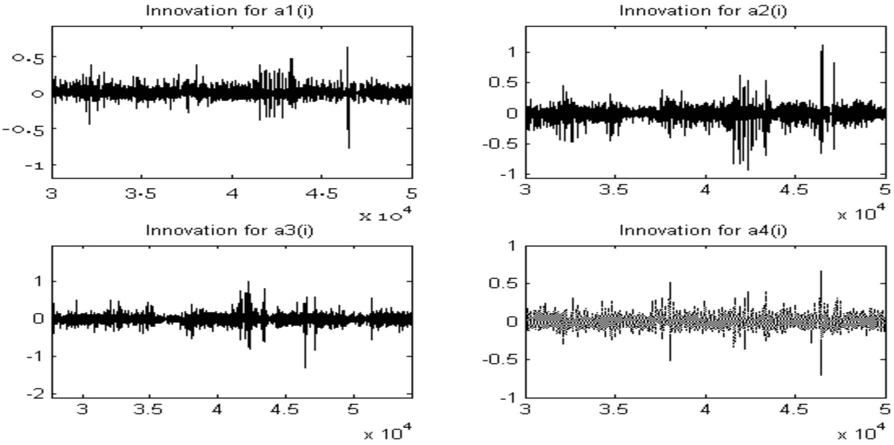


Fig. 6. Innovation courses of four autoregressive parameters

Autocorrelation coefficients were used for estimation of the parameters $c_{j,k}$, as well as the λ_j^2 of the model (6).
In the research, $dC = 3$ was assumed, and coefficients $c_{j,k}$ were calculated with the use of a speed and simple algorithm given in [10], where $R_{\Delta_j(0)} \equiv r0j$

```
% Initial values for j=1:
C11(1) = 0; C12(1) = 0; C13(1) = 0;
L1(1) = r01;
%
    C13(m)=r31/L1(m-1);
    C12(m)=r21/L1(m-1)-C11(m-1)*C12(m-1);
    C11(m)=r11/L1(m-1)-C11(m-1)*C12(m-1)-C12(m-1)*C13(m-1);
    L1(m)=r01/((1+C11(m))*(1+C11(m))+C12(m)*C12(m)+C13(m)*C13(m));
end
```

In this research, every potential speaker S_n was characterized by a feature matrix

$$\mathbf{M}_n = [\mathbf{f}_1, \mathbf{f}_2, \ldots \mathbf{f}_m]$$

where $\mathbf{f}_k$ for $k = 1 \ldots m$ is a vector of features estimated from the $k - th$ utterance of the speaker S_n. Dimension of the $\mathbf{f}_k$ results from the assumed model orders, $dA = 4$ and $dC = 3$, and m=5 different utterances of each speaker were considered. Therefore,

$$\mathbf{M}_n = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,5} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,5} \\ c_{3,1} & c_{3,2} & \cdots & c_{3,5} \\ \lambda_{1,1} & \lambda_{1,2} & \cdots & \lambda_{1,5} \\ \cdots & \cdots & \cdots & \cdots \\ c_{1,4} & c_{1,4} & \cdots & c_{1,5} \\ c_{2,4} & c_{2,4} & \cdots & c_{2,5} \\ c_{3,4} & c_{3,4} & \cdots & c_{3,5} \\ \lambda_{4,1} & \lambda_{4,2} & \cdots & \lambda_{4,5} \end{bmatrix}$$

Features of N potential speakers S_n were gathered in a speaker base $\mathbf{B}$.

$$\mathbf{B} = [\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n]$$

Tested speaker S_x was identified on the base of one utterance only, different from the utterances used in speaker base, and was characterized by $\mathbf{f}_x$. K-means algorithm was used for classification. Up to now, small speaker bases (N=5) were tested, and the method appears to be fast and accurate in 90%.

5 Summary

Results of the research conducted up to now point out that bilinear approximation of non-stationary autoregressive systems may be considered as a tool for non-stationary systems classification. The method is simple and fast e.g. in comparison with classic methods applied for speaker recognition. However, it should be tested for bigger base of speakers. The key problem seems to be the algorithm of MA model identifications, because it was observed that for a few realizations of the same process, the method was unstable, i.e. estimated c_j parameters tended to infinity. Such runs of $y(i)$ were replaced with another realization of the same process, giving stable solution.

References

1. Bensty, J., Sondhi, M., Huang, Y. (eds.): Springer Handbook of Speech Processing. Springer (2007)
2. Box, G., Jenkins, G.: Time series analysis forecasting and control. Holden Day (1970)

3. Campbell, J.: Speaker recognition: A Tutorial. Proceeding of the IEEE 85(9) (1977)
4. Wang, X., Zhang, X., He, Z.: Adaptive algorithm for consistent MA parameter estimation via third order cumulant. Journal of Electronics 14(2), 159–164 (1997)
5. Dimitriou-Fakalou, C.: Yule-Walker Estimation for the Moving-Average Model. International Journal of Stochastic Analysis 2011, Article ID 151823, 20 pages (2011) doi:10.1155/2011/151823
6. Kohlmorgen, J., Lemm, S.: An On-lLine method for segmentation and identification of non-stationary time series. In: Proceeding of Neural Networks for Signal Processing XI, pp. 113–122 (2001)
7. Ludwig, M.: Building on Durbin's method to estimate MA processe. Improving Durbin's method to estimate MA processes arXiv:1304.7956 [stat.ME], `http://mludwig.org/research.html`
8. Ni, H., Yin, H.: Self-organising mixture autoregressive model for non-stationary time series modelling. Int. Journal of Neural Systems 18(6), 469–480 (2008)
9. Ozaki, T., Tong, H.: On moving average parameter estimation. In: Proceedings of the 8th Hawaii International Conference on System Science, pp. 224–226 (1995)
10. Pollock, D.S.G.: A Handbook of time series analysis, signal processing and dynamics. Academic Press (1999)
11. Sandgren, N., Stoica, P., Babu, P.: On moving average parameter estimation. In: Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2348–2351 (2012)

Enhancements of Moving Trend Based Filters Aimed at Time Series Prediction

Jan Tadeusz Duda and Tomasz Pełech-Pilichowski

AGH University of Science and Technology,
Faculty of Management, Department of Applied Computer Science,
Al. A. Mickiewicza 30, 30-059 Krakow
{jdu,tpp}@agh.edu.pl

Abstract. One of the techniques recommended for calculating the nonparametric trend (nonstationary, low-frequency time series component) is the moving trend based smoothing [10], [3], [4] typically based on linear Least Square (LS) approximates of the series in a moving window. Such algorithm splits input time series into three sections, starting, central, final ones. The procedures used in each section may be viewed as Moving Trend based Filters (MTF). In the paper MTFs properties in frequency domain are considered, from seasonal time series decomposition, smoothing and prediction efficiency perspectives. A number of MTFs enhancements is proposed, involving different approximating polynomials and final section signal corrections. In particular, a compression of the final section MTF output signal is proposed to reduce the filter delay, and then reconstruction of the missing signal by a special multiple LS approximation. It improves significantly the final section signal shape evaluation, which are essential for further prediction purposes.

Keywords: Moving trend filters, frequency domain analysis, time series prediction.

1 Introduction

Nonstationary time series processing may be aimed at low-frequency component separation (detrending) by the series smoothing [9] and then – analyzing the properties of periodic and stochastic components, to facilitate further reliable analysis such as time series prediction, event detection, pattern recognition, computational intelligence algorithm implementation [6]. The nonparametric long-term trend may be drawn from a nonstationary time series by the series filtering with moving trends [8], [3], [2], [4]. Time series filtering procedures, i.e. moving average filtering, exponential smoothing, Hodrick-Prescott (H-P) filtering, mostly based on linear filtering methods, are used in many fields [9].

The classical moving trend filter (MTF) is based on the rolling approximation of the processed time series with the least-square (LS) linear approximation in a moving window [8], [3], and then averaging of the approximates found for consecutive time points [2], [6]. Such procedures splits the series into three sections: the starting (s),

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_71, © Springer International Publishing Switzerland 2014

central (c), final (f) ones. Having a given window width (M), the starting section begins at the first (oldest) sample and finishes with the (M-1)-th one, the filtering window width enlarges from M to 2M-2, and the number L_n of the approximates $y_F(i, t_n)$ to be averaged increases from 1 do the M-1. The central (c) section ranges from the M-th to $n - M+1$ samples, the filtering window width is 2M–1, and the number of approximates is M. The final (f) section contains the samples from n – M + 2, to the newest one *n-th*, the window width reduces from 2M-2 to M, and the number of approximates $y_F(i, t_n)$ to be averaged decreases from M-1 to 1 [4], [5].

The moving window width (M) is adjusted in such a way to reach appropriate smoothing of the series and periodic component extraction in the central section. The window width affects the trend properties and cyclic components separation effectiveness [4]. However, in our papers [5], [4] we have shown that the final section filters strongly distort the trend, which strongly decreases a usefulness of the classical MTF for prediction purposes. We have also shown that much better smoothing and prediction properties can be reached by employing in MTF higher order polynomial approximations, and by specification of the required filter properties in frequency domain (a number of the filter variants have been examined by analysis of Bode plots [11]). In particular, one and two-parameter filters have been discussed and finally recommended [4].

In this paper certain essential improvements of MTFs are proposed, aimed at time series prediction enhancement, especially addressed to seasonal series. In particular, a compression of the final section MTF output signal is proposed to reduce the filter delay, and then reconstruction of the missing last section signal by a special multiple LS approximation. Effects of the corrections are shown on an example of a noised periodic signal processing results.

2 Nonstationary Time Series Processing with MTF Filters

Nonstationary time series *y(t)* (see eq.1) may be considered as the sum of an aperiodic trend function *f(t)*, a cyclic component *C(t)* of time period *T*, and a higher frequency zero-average noise *z(t)* [1], [7], [5], [4].

$$y(t) = f(t) + C(t) + z(t)\,. \tag{1}$$

One assumes the periodic component (see eq.2 where A_k denotes the amplitude of i_k-th harmonic, τ_k is the delay of the k-th component) consists of a number of harmonics specified by a set of indices i_k, $k = 1, \ldots K$ (e.g. $i_{k} = 1, 2, 10$) representing some known seasonal phenomena affecting the series. It can be expressed as follows [11], [5], [4]:

$$C(t) = \sum_{k=1}^{K} A_k \sin\left(\omega_T i_k (t - \tau_k)\right), \quad \omega_T = \frac{2\pi}{T}\,. \tag{2}$$

The nonparametric trend *f(t)* may be calculated for each time step t_m with a low-pass digital filter MTF adapted to extraction of high frequency components (ω_T and higher) from the original (input) diagnostic signal. The *C(t)* component can be removed from the filtering residuals by the LS approximation, i.e. by identification of the regression model of the form (2) [1], [5],[4]. The $f(t_m)$ values are calculated by averaging all approximates $y_F(m,k)$ of the series found in a moving windows containing M samples and covering the m-th sample.

The Y_{Fk} approximates in a window ending at (n-k)-th sample (n denotes the most recent value) are calculated with the following LS formula:

$$\begin{gathered} Y_{Fk} = Y_k \cdot W, \\ W \overset{def}{=} U[U^T U]^{-1} U^T, \quad u_i = [1, t_i^{p1}, \ldots t_i^{pL}], \ \cdot \\ i = 1, \ldots, M, \qquad t_i = -M+1, \ldots, 0 \end{gathered} \tag{3}$$

where $Y_k = [y_{n\text{-}k\text{-}M+1}, \ldots, y_{n\text{-}k}]$, $Y_F = [y_F(t_1), \ldots, y_F\ (t_M)]$, U denotes the model input matrix consisting of M rows $u_i(t_i)$, t^{pk} denotes selected p_k-th order monomials, W is a constant M×M filtering matrix.

The series $f(t_m)$ found in this way may be further used to h-samples ahead prediction of the series main component $f(t_n + h)$, by its extrapolation with a h-samples increment $\Delta_h f$ averaged with harmonic weights [3], [5], as shown in equations (3) [4]:

$$\begin{gathered} f(t_{n+h}) = f(t_n) + \Delta_h f, \ \ \Delta_h f = \sum_{i=1}^{n-h} C_i \big(f(t_{i+h}) - f(t_i)\big), \\ C_0 = 0, \ \ C_i = C_{i-1} + \frac{1}{(n-h)(n-h-i+1)} \end{gathered} \ . \tag{4}$$

Considering the classical algorithm (labeled as z1) the approximation formula is calculated with the following equation: $y_F(t_i) = b_0 + b_1 t_i$, i.e. u_i=[1, t_i], while for the moving average (labeled as z0) the formula is $y_F(t_i) = b_0$, i.e. u_i=[1].

3 MTF Filter Generalization

In our earlier papers [5], [4] we have proposed a generalization of the moving trend smoothing algorithm, by employing higher order approximating polynomials of appropriately designed properties. In particular, the following formula (labeled as s3) has been applied and described:

$$y_F(t_i) = b_0 + b_3 t_i^3, \text{ i.e.. } u_i = [1, t_i^3]. \tag{5}$$

The impulse response coefficients $g_{\{s,\ c,\ f\}}()$ of the moving trend filters (MTF) attributed to the three sections (starting, central, final ones) have to be calculated by summing the selected elements of the W matrix to give M-1 g_s filters for the s (starting) section (ranged between M and 2M-2), M-1 g_f filters for the f (final) section (of length between 2M-2 and M), and g_c filter for the central c section (length: 2M-1).

By applying the Fourier Transform to the filters g_s, g_c, g_f one may examine properties of different MTF filters (involving different approximating polynomial types) in frequency domain. The dependencies between the signal harmonics related to frequency (gain diagrams end delay profiles – see fig. 1 and 2) allow predicting the filter properties in time domain (such as the periodic component separation efficiency, the nonparametric trend sensitivity to noise content, and the trend distortion level).

The most efficient separation of the periodic component (of a period T) is obtained when the first low-frequency gain minimum is at ω_T point. In our papers [5],[4] we have shown that the minimum position depends on the M/T ratio. For the filters z1 and s3 this minimum occurs at the ratio M/T=1.38 (as shown in fig. 1) while for the z0 filter the ratio is to be fixed at M/T=1.

Fig. 1 and 2 show a comparison of frequency properties of the z0, z1 and s3 filters of the central and final section, related to the 4th order recursive Butteworth low-pass filter [10] (labeled as B4), having at ω_T the same gain as MTF of s3 type (with a half-gain frequency equal to ω_T*0.55). The s3 filter of the central section (bold line) is a bit worse than z1 (faster gain decrease in the pass-band, a bit greater gain in the attenuation band). The s3 advantages become evident, when looking at the gain profiles of the final section filter, which heavily affects the predictor properties. It is due to the much smother amplitude characteristics in the pass-band, the much smaller spectral leakage in the attenuation band (see fig. 1) and advantageous delay properties in the pass-band (see fig. 2). The s3 filters have less varying delay profiles that z1 that announces less trend distortion. Note that the B4 delays are unacceptably large and highly varying, thus this filter is practically unusable for prediction purposes.

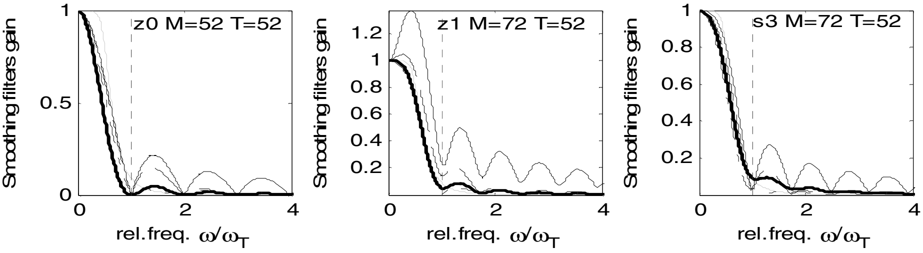


Fig. 1. Gain diagrams (v.s. $\omega/\omega T$) for the smoothing filters z0, z1 and s3 – central section: h<-M (bold lines), the final section: h=0 (solid lines, causal filter), h= -24 (dotted lines), h=-47 (point-dotted lines), shadow solid lines (B4)

Frequency properties presented above (figures 1 and 2) are visible in time domain responses. Fig. 3 shows effects of a periodic time series filtering with z0, z1 and s3 filters. The z0 and s3 filters produces much smaller step-size overestimation than the z1 and much shorter response time than the Butterworth's (B4) filter (see fig.3).

Figure 4 illustrates the quality of the periodic signal smoothing in the central section (at the left of the vertical dashed line) and in the final section (at the right).

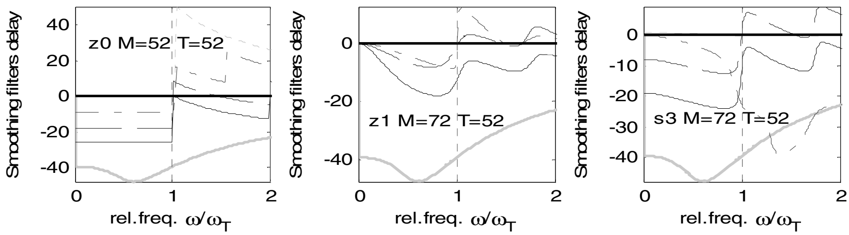


Fig. 2. Delay profiles of the final section filters z1 and s3 – the final section: (*h*=0 solid lines), *h*= -24 (dotted lines), *h*=-47 (point-dotted lines), *h*=-60 (point lines), *h*=-71 (solid lines close to the zero delay), bold-line - for 0 delay of central section filter; shadow lines – delay of the B4

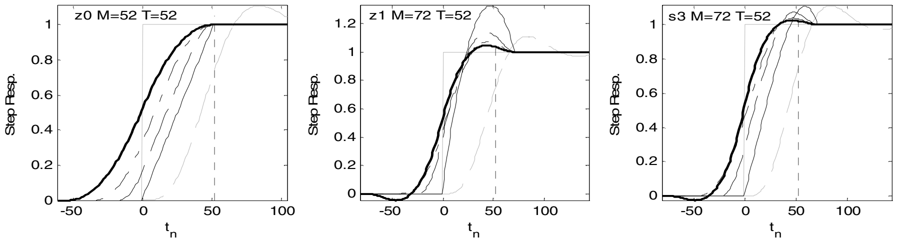


Fig. 3. Step-wise filter responses for the signal *y*(*t*) (shadow bold line): central segment filter (bold line, *h*<-*M*), final one (final filter response *h*=0: solid line, *h*=–40: dotted line, *h*=–-20: dashed line), step response of the reference B4 low-pass filter (shadow line)

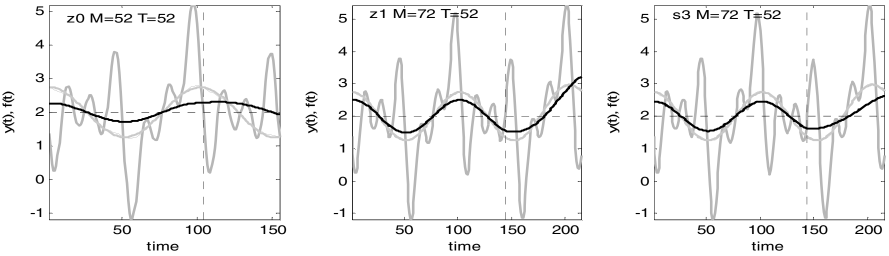


Fig. 4. The quality of the periodic signal filtration (shadow line) in the central section (to the left of the vertical dashed line) and in the final section (to the right). Signal to be restored: wavy blue line, signal approximated in the final section (original one after the cyclic component extraction by regression analysis for the central section): shadow line (rapidly changing): filtered signal, noiseless one: bolded line.

For the classical filter z1 an excessive trend gain in the final section is clearly seen. The signal distortion is due to the delay profile irregularity in the whole band, the excessive gain in the pass-band and the significant spectral leakage in the attenuation band (see fig. 1).

4 Proposed MTF Enhancement

The amplitude/delay profiles shown in the fig. 1 and 2 suggest that at least for the z0 and s3 filters of the final section the signal distortion may be significantly reduced through the signal compression by the pass-band averaged delay value (the z0 and s3 filters delay is relatively constant, while the gain profile is similar to that of the central section filter)

Fig. 5 shows the averaged delay curves of the final section filters. The averaging is made over the pass band with the frequency dependent weights $\omega A(\omega)/s_\omega$, $\omega \in (0, \omega_T)$, where s_ω denotes the sum of the expressions in the numerator.

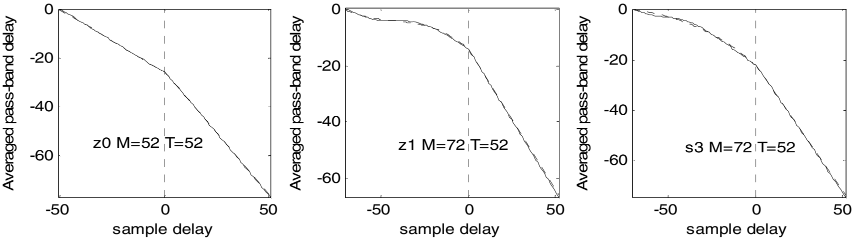


Fig. 5. The final section filters delay (averaged over the pass-band), versus the sample position related to the final section end (the last sample has 0-position, $i \leq 0$ position denotes consecutive points obtained by smoothing with the $g_f(M\text{-}i\text{-}1)$ filters, $i>0$ position: prediction with extrapolation of the increments averaged by the harmonic weights [2]. Dashed lines – the final filters delay approximation ($i \leq 0$) and $\tau_i = i$ line for predictors ($i>0$).

In our research we have stated that the final section delay curves shown in figure 5 ($i \leq 0$) may be well approximated by the following formula (eq. 6):

$$\eta_i \overset{def}{=} \frac{\tau_i}{M} = \left(1+\frac{i}{M-1}\right)\left(\frac{\tau_0}{M} + D_1\frac{i}{M-1} + D_2\frac{i}{M-1}\left(\frac{i}{M-1}-1\right)\right), \qquad (6)$$
$$i = -M+1, \ldots.,0\,.$$

where the coefficients D_1 and D_2 are determined with LS method for the fixed values $\tau_{1\text{-}M}=0$ and $\eta_0 = \tau_0/M$ (according to fig.5).

By changing the positions of subsequent filters g_{fi}, $i = -M+1, \ldots., 0$, to the positions $j = M\tau_i$ (see eq.6), with canceling the filters previously set at this point, we obtain a revised package of the g_{fsj} filters of the same amplitude characteristics as g_{fi}, but of reduced delays. Delay profiles versus relative frequencies are shown in the fig. 6.

Notice the rapid delay change at the ω_T point, especially for the z0 and s3 filters whose amplitude properties are better than z1's (see fig. 1). Therefore, to avoid the compressed signal distortion, before the final filtration of seasonal time series, we recommend to remove the periodic component from the original series in the final section. It can be done by identifying the regression model of the form (2) with the central section MTF filtering Least-Squares fitting.

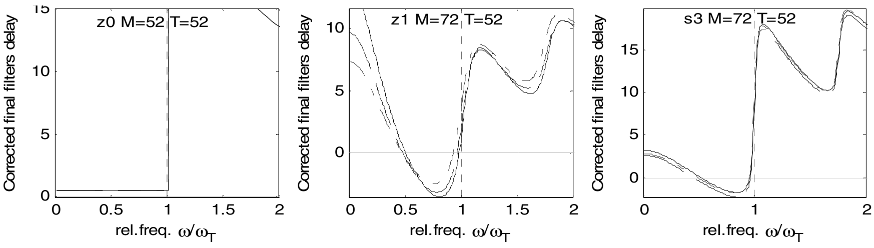


Fig. 6. Delay profiles versus relative frequencies diagrams for the z0, z1 and s3 filters of the final section after the g_f final filters shift (eq. 5) for h=0 (solid line), h=-24 (blue dashed line), h=-47 (red dotted line), h=-60 (green dotted line), h= 71 (solid line near to zero delay), bold line – for the zero-delay filter of the final section, related to the B4 filter. Nonparametric trend: $f(t)$=sin(2π /($2T_u$)·t)·0.85-sin($2\pi t$/($3T_u$)·t·0.25+sin($2\pi t$/($4T_u$)·t)·0.15 Periodic component: $C(t)$=sin($2\pi/T_u$·t)-sin($4\pi/T_u$·t)+sin($6\pi t/T_u$·t).

The results of the proposed procedure are illustrated in the fig.3. Signals get with original series filtering (solid line) are compared to the signal after the periodic component extraction. Effects of the signal compression (with g_{fs} filters) are shown in the fig. 6, together with further processing results.

To avoid distortion resulting from varying delay of high-frequency random components, one may recommend the additional filtration of compressed subseries by the moving average in the window of the length L_f=2·d_f+1, where d_f=max{int(T/50), 1}. It was found that the L_f value should be ranged between 3 and 11 samples. Thus the compression and filtering of useful signal cause loss of samples of the final section ranged from j_{A0}=$M\eta_0$ – d_f +1 sample to j_0=0 (the newest one), L_A=–j_{A0}+1 samples in total. For prediction efficiency purposes, the essential is reliable reconstruction of the loss samples, especially the last-sample-estimation.

We have stated that good estimates may be produced through multi-variant fitting of suitably smoothed signal profile to the last L_A data. It should be done in three stages:

a) As an approximating function $f_A(t_m)$, the 4rd order polynomial is assumed, having zero derivative at the end of the fitting interval and three fixed values over the compressed signal section (m<j_{A0}), equal to the compressed signal values $y_{Ffs}(t_m)$: $f_A(t_{(-LA-k)})$= $y_{Ffs}(t_{(-LA-k)})$, k={0, 1, d_2}. By the samples of k=0 and k=1 an approximation smoothness is ensured, while d_2 is fixed by trials.

b) The conditions imposed on the approximate bind four parameters a_0, a_1 a_2 and a_4 of the polynomial with the fitting parameter a_3 (see eq. 7),

$$a_0 = y_{Ffs}(t_{-LA}), \quad a_1 = a_{10} + B_1 a_3, \quad a_2 = a_{20} + B_2 a_3, \quad a_4 = a_{40} + B_3 a_3$$

$$\begin{bmatrix} a_{10} \\ a_{20} \\ a_{40} \end{bmatrix} = A^{-1} \begin{bmatrix} y_{Ffs}(t_{-L_A-d_2}) - y_{Ffs}(t_{-L_A}) \\ y_{Ffs}(t_{-L_A-d_2}) - y_{Ffs}(t_{-L_A}) \\ 0 \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = -A^{-1} \begin{bmatrix} (-d_1)^3 \\ (-d_2)^3 \\ 3L_A^2 \end{bmatrix}, A = \begin{bmatrix} -d_1 & d_1^{\ 2} & d_1^{\ 4} \\ -d_2 & d_2^{\ 2} & d_2^{\ 4} \\ 1 & 2L_A & 4L_A^3 \end{bmatrix} \qquad (7)$$

The model f_A is fitted with a parameter a_3, using the generalized LS method

$$a_3 = [Y_R - y_{Ffs}(t_{-LA}) - a_{10}t_R - a_{20}t_R^2 - a_{40}t_R^4]\cdot \varphi_R^T[\varphi_R\varphi_R^T]^{-1}, \tag{8}$$

$$Y_R = y(t_{-LA+t_R}),\quad t_R = [1,...,L_A],\qquad \varphi_R \overset{def}{=} B_1t_R + B_2t_R^2 + B_3t_R^4 + t_R^3,$$

The values for $f_A()$ over the fitting interval t_R are calculated as follows:

$$f_A(t_{-L_A+t_R}) = y_{Ffs}(t_{-LA}) + (a_{10} + a_3B_1)t_R + (a_{20} + a_3B_2)t_R^2 + a_3t_R^3 + (a_{40} + a_3B_3)t \tag{9}$$

c) The value for d_2 is selected by trials, to obtain the $f_A(t_m)$ approximate for $m=-L_A$, …,0 of similar shape to the signal $y_{Ffs}(t_m)$ at the beginning section of the final segment, i.e. for $m=-M$, .., j_{A0}-1. We have found that the most appropriate similarity measure is the difference module of the second increments mean squares of the series $f_A(t_i)$ and $y_{Ffs}(t_j)$.

Equations (8) and (9) can be written as the G_{fA} matrices of digital filters which joint to the corrected filters matrices g_{fs} gives the full matrix G_{fs} of modified FIR filters for the final segment.

Efficiency of the proposed filtering method in the final MTF segment for the periodic series (see fig. 3) is illustrated in the fig. 7.

Fig. 7 shows that the proposed method (especially combined with the s3 filter) allows for significantly enhancements of the final signal value estimation in much more difficult case, i.e. for highly noised nonstationary signal containing additional harmonic of period $2T/3$ and amplitude equal to 1 (almost the same as other harmonics in total). For all examined filters, including the B4 recursive filter, this harmonic is passed weaker than lower frequency components (see fig. 3 and 6). It strongly disturbs the filtering residuals in the central segment, thus makes difficult to identify A_k and τ_k parameters of the periodic component (2). Useful signal was distorted by a strong autoregressive process of standard deviation 0.5. In addition, to force the reconstructed signal nonstationarity, the signal was disrupted by the mean value step changes of two units for three subsequently time instants.. The filter responses for three described above cases (in three subsequent rows) are shown in the fig. 8. Fig. 8 shows that in all cases, a satisfactory accuracy of the trend (signal) estimation in the final section is reached, much better than without the proposed enhancement. Especially good results are produced by the s3 filter - the approximates reach the most similar shape to the reconstructed signal. It corresponds to our expectations based on frequency properties illustrated in fig. 1 and 2.

Fig. 6 and 7 show large diversity of the selected (the most advantageous) d_2 value. It confirms a validity of the proposed idea focused on approximate properties adaptation to temporary spectral properties of the reconstructed signal in the final section.

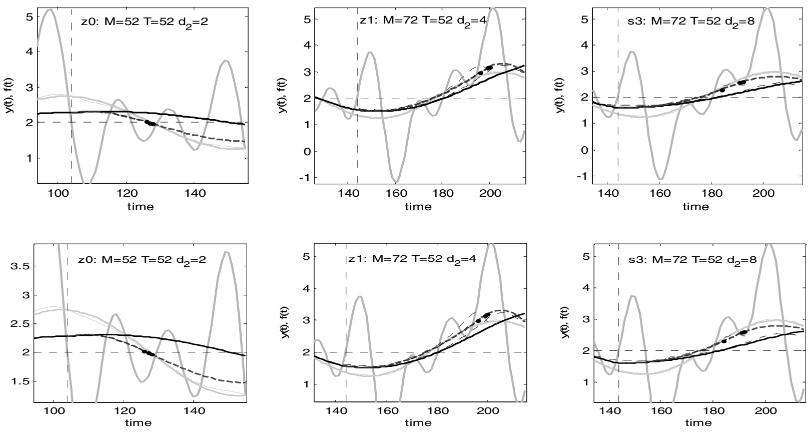


Fig. 7. Effects of the useful signal reconstruction (see fig. 3) in the final section with the proposed method. Black dots on the red line denotes the approximate anchorages.

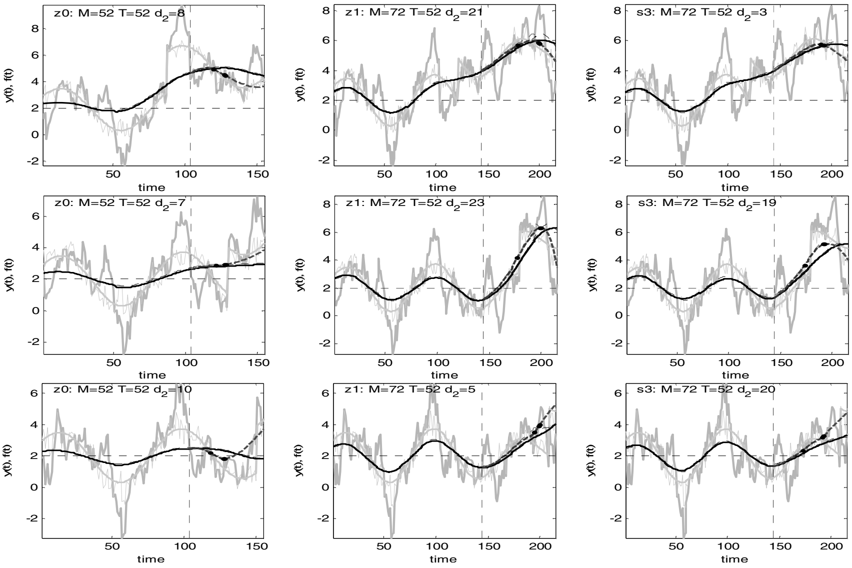


Fig. 8. The quality of nonstationary periodic signal filtering (green color) of the central section (at left of the dashed vertical line) and the final section (at right). Total reconstructed signal: solid cyan line (including random noise drawn with the same color); reconstructed signal with MTF: bold line; compressed signal in the final section: red dashed line; reconstructed signal with approximation: magenta dashed line; step-change was added 24 samples (in the top subfigures), 36 samples (in the middle figures) and 8 samples (in the bottom subfigures) before the final section beginning. Periodic component: $C(t)$=sin($2\pi/T_u \cdot t$)-sin($4\pi/T_u \cdot t$)+sin($6\pi/T_u \cdot t$) Random noise: $z(t_n)=\alpha \cdot z(t_n\text{-}1)+0.5\cdot(1\text{-}\alpha^2)^{1/2}$, α=0.15.

5 Conclusion

The proposed MTF improvement (especially addressed to seasonal series) is based on the useful signal (low frequency component) correction in the series final section. To this aim, the three-stage-procedure is proposed: (a) extracting the periodic component $C(t)$ from the original series in last section (this component may be identified by a linear regression method applied to the MTF filtering residuals in the central section); (b) applying the MTF filters for the signal obtained in the step (a) in the last section, and then, in order to reduce the delay effect of the final section filters, compressing the filtering output by shifting samples from time instants k_{fi} =1, …, M-1 to the positions k_{fsi}= k_{fi} -(M-1)·h_i, 0≤h_i <0.5 (the obtained signal may be smoothed with moving average in the window of width L_f=2·d_f+1, d_f=max{int(T/50), 1}, L_f should be from 3 and 11 samples); (c) completing the missed trend samples in the range <$k_{fs(M\text{-}1)}$-d_f+1, M−1> by fitting an appropriate 3th order polynomial to the original series, assuming a shape similarity of the produced signal to the signal reached at the stage (b).

The difference module of the second increments mean squares of the series found at stage (b) and (c) may be taken as the similarity measure at the stage (c). The results shown in fig. 7 and fig. 8 confirm a validity of the proposed approach. The last section MTFs compression at the stage (b) reduces the final filter delay and reconstruction of the missing last section signal by the fitting procedure at the stage (c) improve significantly the final section signal evaluation, which is essential for further prediction purposes.

References

1. Askom, M.V., Chenouri, S., Mahmoodabadi, A.K.: ARCH and GARCH models. Dept. of Statistics & Actuarial Sciences, University of Waterloo (2001)
2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis, Forecasting and Control, 3rd edn. Prentice Hall, Englewood Cliffs (1994)
3. Cieślak, M. (ed.): Prognozowanie gospodarcze. Metody i zastosowanie, PWN Warszawa (2002)
4. Duda, J.T., Pełech-Pilichowski, T.: Moving Trend Based Filters and Predictors Properties in Frequency and Time Domain. In: 6th EuroSymposium on Systems Analysis and Design. Gdansk, Poland (in press, 2013)
5. Duda, J.T., Pełech-Pilichowski, T.: Moving Trend Based Filters Design in Frequency Domain. In: Proc. of the Jubilee XX International Symposium on Application of Systems Theory. Automatica. AGH UST University Press (in print, 2013)
6. Duda, J.T., Pełech-Pilichowski, T.: Opracowywanie prognoz sytuacji hydrogeologicznej i ostrzeżeń przed niebezpiecznymi zjawiskami zachodzącymi w strefach zasilania lub poboru wód podziemnych. Research Report, AGH UST Faculty of Management, Kraków (2012)
7. Hamilton, J.: Time Series Analysis. Princeton University Press (1994)
8. Hellwig, Z.: Schemat budowy prognozy statystycznej metodą wag harmonicznych. Przegląd Statystyczny (2) (1967)
9. Kim, S.J., Koh, K., Boyd, S., Gorinevsky, D.: L1 trend filtering. SIAM Review (2009)
10. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing. Prentice-Hall, Englewood Cliffs (1989)
11. Otnes, R.K., Enochson, L.: Digital Time Series Analysis. John Wiley, New York (1972)

Indivertible Elementary Bilinear Time-Series Models for Data Encryption

Łukasz Maliński

Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
lukasz.malinski@polsl.pl

Abstract. It has been shown that coefficients of indivertible elementary bilinear time-series models can be estimated with almost no bias. Moreover, the random processes obtained by simulation of such models are not correlated. Therefore, those features suggest a possibility of practical application of indivertible elementary time-series models for data encryption.

Keywords: Bilinear models, indivertible models, data encryption, time-series.

1 Introduction

Bilinear time-series models are one of the simplest examples of nonlinear time-series models, and they are a natural extension to alreadywidelyexploited linear time-series models, which are the most popular up to this date, used in the field of process identification.

Granger and Andersen [1] are recognised among the pioneers in the field of nonlinear time-series modelling. They are also among the first (1978) authors who placed their interest in bilinear time-series models [2]. Further contribution to this field was made by Subba T. Rao [3] and Quinn [4] in early eighties of previous century. Later in 1987 Gooijger and Heuts [5] have performed an analysis of higher order statistical models for bilinear time-series, building the fundaments for method of moments (MM), which can be used for estimation of model coefficient values, unfortunately with limited efficiency (biased estimates). Two years later Guegan and Pham [6] considered using another popular estimation algorithm called Least Squares (LS) algorithm. They also showed the limited efficiency of this solution.

In 1994, a stability condition for general bilinear time-series model was presented by Lee and Mathews [7] and the same year Bielińska and Nabagło proposed the modification to LS algorithm, which has stabilised an estimation procedure [8] and also reduced the bias of estimates. One year later, Brunner and Hess [9] analysed the solution space for Maximum Likelihood (ML) algorithm used in identification of simple linear-bilinear model. They simulated the bilinear model and showed that the shape of the cost function used by ML algorithm became more complex for larger values of coefficients (near stability threshold). This problem can be denoted as one of the major source of difficulties in identification of bilinear models using ML and LS algorithms [10]. Another source of difficulties was presented in 2011 by Maliński

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_72, © Springer International Publishing Switzerland 2014

and Figwer [13]. Authors showed that computation of the statistical moments for processes obtained from elementary bilinear models can be significantly biased for model coefficients values near stability threshold. This way, the previously shown problem with using MM approach to estimation has been explained.

The modification of LS algorithm proposed in [8] was later (2011) exploited by Maliński [11] to indicate that it is a key to obtain unbiased estimates of coefficient of indivertible elementary bilinear time-series model. Further improvement of this modification, considering it parameterisation was done in 2012, allowing unbiased estimation of coefficient of any stable elementary bilinear time-series model [12].

Moreover, it was shown in [12] that estimates of coefficients obtained for indivertible elementary bilinear time-series model has very low random scatter. This conclusion and correlation properties of bilinear process itself, have brought up the idea to use this models for data encryption. Next sections of this paper contain necessary theoretical background on elementary bilinear time-series models and examples of simple data encryption. Some statistical research has also been performed to analyse an efficiency of the proposed application.

2 Theoretical Background

The general bilinear time-series model, BARMA(dA, dC, dK, dL) model is defined as follows [3]:

$$y(t) = \sum_{i=1}^{dA} a_i y(t-i) + \sum_{j=0}^{dC} c_j e(t-j) + \sum_{k=1}^{dK} \sum_{l=1}^{dL} \beta_{kl} e(t-k) y(t-l) \cdot \tag{1}$$

where: $y(t)$ is an output sequence, t is a discrete time indicator, coefficients a_i and c_j determine linear part of the model, $\beta_{kl}$are coefficients of the bilinear part, dA, dC, dK and dL are structure parameters defining the particular ranks of each model component, $e(t)$ is a stimulation signal assumed to be independent, identically distributed random sequence.

Numerous difficulties, related to identification and analysis of BARMA model, can be found. Therefore, many authors perform their research on simplified bilinear model structures [4-6] and [8-9]. Also, in this paper, the simplest structure - the elementary bilinear time-series model (EB) - is considered. The EB model (2) contains a single component of bilinear part (from BARMA model) and a stimulation sequence $e(t)$, only.

$$y(t) = e(t) + \beta e(t-k) y(t-l) \,. \tag{2}$$

There are some assumptions, related to a stimulation sequence $e(t)$, that have to be undertaken in order to perform the more thorough analysis of EB model statistical properties. In this paper it is assumed that $e(t)$ is white noise of a Gaussian distribution of following statistical properties:

$$E\{e(t)\} = 0; \quad E\{e(t)^2\} = \lambda^2; \quad E\{e(t)e(t-1)\} = 0; \quad E\{e(t)^3\} = 0 \,. \tag{3}$$

This assumption allows to define the stability and invertibility conditions, respectively by (4) and (5):

$$\beta^2\lambda^2 < 1; \tag{4}$$

$$\beta^2 m'^{(2)}_y < 1. \tag{5}$$

where $m'^{(2)}_y = E\{y(t)^2\}$.

The invertibility of EB model, restricted to condition (5) is an important issue in estimation of the EB model coefficient value. Because, $(m'^{(2)}_y \geq \lambda^2)$, the invertibility threshold for EB model is always lower than its stability threshold. Therefore, it is possible to obtain a stable realisation of process generated by EB model which is indivertible (unstable inverse model). Typically, indivertible models are considered to be unidentifiable, but for EB model instability (represented by occasional explosions in output sequence) is not cumulative and can be easily controlled by simple saturation function.

This solution has been presented in [11-12] and will not be further considered in this paper. What is important, the coefficient of indivertible EB model can be estimated with no bias and with very low or even negligible random error so it can be used in data encryption applications.

3 Example of Data Encryption

At the beginning of this section it must be highlighted that data encryption technique presented here is very simple and dedicated for presentation of possibilities only. The data encryption is based on simulation of random process using EB model and the main principle of data coding is to assign a narrow range of model coefficient β values to the text character. During simulation a mean value of the range corresponding to the coded character is used as EB model coefficient and *N* number of random process samples is obtained. The *N* parameter will be called as samples per character.

For the sentence consisting n characters, the sequence of $n \times N$ samples is generated and transferred digitally as encrypted data to receiver of the sentence. The decoding procedure requires:

- the number per character(*N*) value.
- the coding table which contains the ranges of the EB model coefficient values assigned to characters.
- structure parameters (*k* and *l*) of the EB model used for encryption.

The entire encrypted sequence of data is splitted into separated *N*-samples subsequences which correspond to different characters. Then, a decryption of the data is performed using identification algorithm [12] which is capable of acquiring the unbiased estimates of EB model coefficient. Finally, the obtained estimates are compared with coding table characters and are assigned to the particular subsequence. This way the original sentence is restored.

As the presented encryption technique seems to be very simple and naïve, the encrypted data even if intercepted will be safe. Because of the fact that they will be probably taken as random noise due to no correlation between their particular values, even though the information provided by correlation will be incorrect. This is presented in Figure 1, where the text “Hi World” is encrypted using proposed technique. The EB model structure was set to $k = 1$ and $l = 2$.

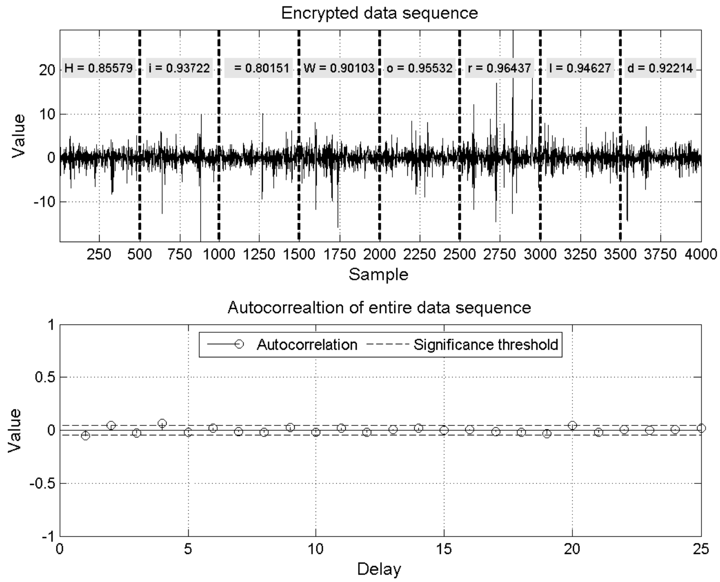


Fig. 1. Autocorrelation for encrypted data

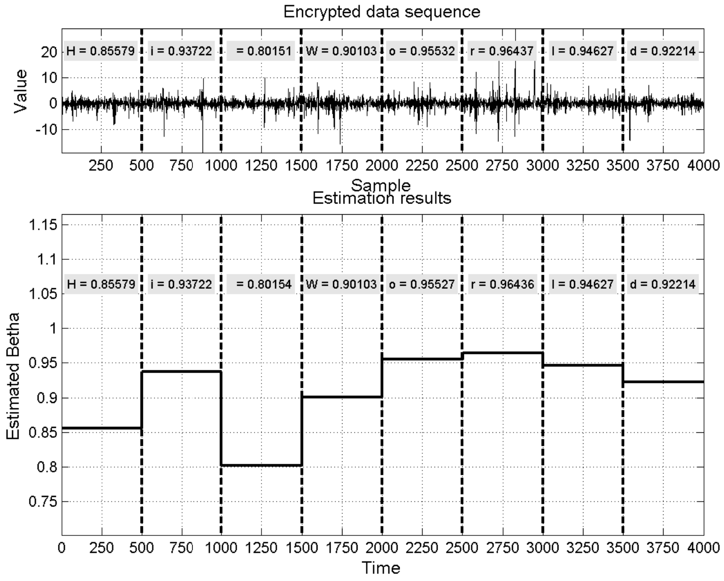


Fig. 2. Results of estimation using EB model of known structure

The top chart represents the encrypted data set, while the bottom chart contains first 25 coefficients of autocorrelation function for entire encrypted data set. The dashed line in the bottom chart represents the significance threshold for coefficients values for autocorrelation.The bottom chart of Figure 2 presents the estimation results of coefficient value of EB model.

It is possible to restore information from the uncorrelated encrypted data with very good precision, what is shown in Figures 1 and 2. Although, the ranges of EB model coefficient are respectively very narrow, there seems to be no problem with restoring original information. In following section, the analysis of effectiveness of data decryption is performed considering a number of samples per character *N* used in data encryption.

4 Results of Experiments

In order to check the impact of *N* value on encryption/decryption efficiency, the simple test wasdesigned. At the beginning, the effective coding range *C* has been set to $\beta \in <0.80, 0.99>$. This range was divided into 63 subranges respectively assigned to following text characters:

- space (first subrange assigned).
- numbers 0,1,2,...,9.
- uppercase letters A,B,...,Z.
- lowercase letters a,b,...,z (last subrange assigned to “z”).

Next, the following phrase: “The indivertible elementary bilinear time series models for data encryption”, which contains $n = 75$ characters, was encrypted in $R = 100$ independent runs (described as r_i for $i= 1, 2, \ldots, R$) of encryption procedure for each *N* value taken from *A* set, where $A=\{50, 100, 250, 500, 750, 1000\}$.

Afterwards, the decryption procedure was performed on each encrypted data sequence. Each decrypted character for particular encrypted sequence r_i was compared to original one. If the result of comparison was an accordance (decrypted character was equal to original one) a Success Counter s_i was incremented by one. If the result of comparison was a divergence (decrypted and original characters did not match) and the identification result was beyond used coding range the decrypted character was marked as unrecognised and the Unrecognised Counter u_i was incremented by one. Finally, if the result of comparison was a divergence, but decrypted character was within coding range (wrong character was decrypted), no counter was updated.

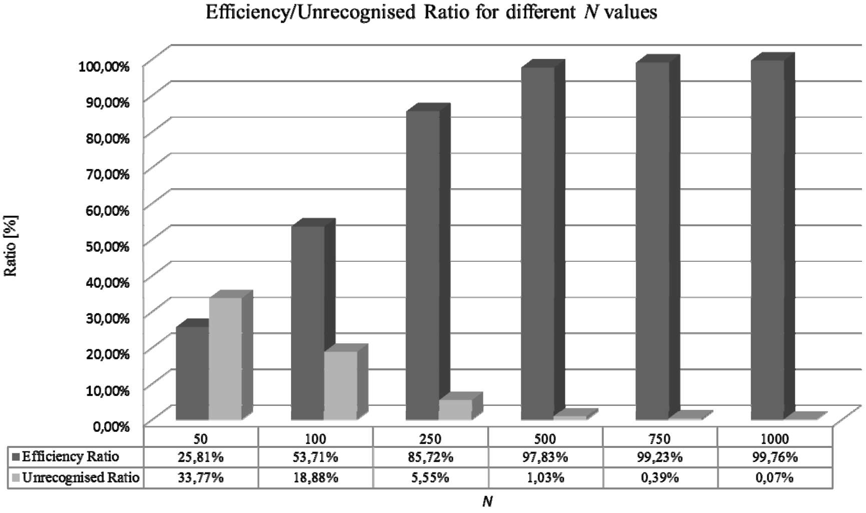
When the Success and Unrecognised Counters were computed for each encrypted sequence r_i, following performance measures were obtained:

- Efficiency Ratio – a percentage of successfully decrypted characters, defined as:

$$ER = \frac{100\%}{Rn} \sum_{i=1}^{R} s_i \, . \tag{6}$$

- Unrecognised Ratio – a percentage of unrecognised characters (identification results beyond used coding range), defined as:

$$UR = \frac{100\%}{Rn}(Rn - \sum_{i=1}^{R} u_i) \cdot \tag{7}$$



	50	100	250	500	750	1000
■ Efficiency Ratio	25,81%	53,71%	85,72%	97,83%	99,23%	99,76%
■ Unrecognised Ratio	33,77%	18,88%	5,55%	1,03%	0,39%	0,07%

Fig. 3. Efficiency and Unrecognised Ratio results

The results obtained for different *N* values, presented in Figure 3, shows that efficiency of the data decryption is strongly *N* depended. For low *N* values the Efficiency Ratio is very low (below 30% for *N* = 50 samples per character) and it reaches almost 100% efficiency for high *N* values (*N* = 1000 samples per character). Also the *N* value influences the number of unrecognised characters. Higher *N* value is followed by more decryption results placed in proper coding range and the remaining divergences are related to small estimation errors resulting in recognition of wrong character.

5 Summary

The final, statistical results of encryption/decryption algorithm efficiency seems to be less spectacular then it was expected from initial analysis. At this stage of development, proposed encryption technique requires a lot of data samples to be efficient. However, it must be stated that data are encrypted by nonlinear, stochastic model and are uncorrelated in its encrypted form, so they might be difficult to be cracked. Moreover, the proposed encryption technique is absent from any correction mechanisms.

The results presented in this paper should be treated as a preliminary results only. This is the first attempt to use indivertible bilinear time-series models in practical application, since those models were assumed to be worthless. There is also a lot of work to be done by studying the sources of errors and it is certain that a lot of improvement is still to be made.

Also, at this point the following improvements are considered for implementation:

- The uniform subranges designated to characters can be replaced by non-uniform ones due to the increasing accuracy of identification for higher β values.
- Some correction mechanisms must be implemented to deal with possible identification errors.

References

1. Granger, C., Andersen, A.: Nonlinear time series modelling Applied Time series analysis. Academic Press (1978)
2. Granger, C., Andersen, A.: An introduction to bilinear time series models. Vandenhoeck and Ruprecht (1978)
3. SubbaRao, T.: On the Theory of Bilinear Time Series Models. Journal of the Royal Statistical Society B44, 244–255 (1981)
4. Quinn, B.: Stationarity and invertibility of simple bilinear models. Stochastic Processes and Their Applications 12, 225–230 (1982)
5. Gooijger, J., Heuts, R.: Higher order moments of bilinear time series processes with symmetrically distributed errors. In: Proceedings to Second International Tempere Conference in Statistics, pp. 467–478 (1987)
6. Guegan, D., Pham, D.T.: A Note on the Estimation of the Parameters of the Diagonal Bilinear Model by Method of Least Squares. Scandinavian Journal of Statistics 16, 129–136 (1989)
7. Lee, J., Mathews, J.: A Stability Condition for Certain Bilinear Systems. Signal Processing 42, 1871–1973 (1994)
8. Bielińska, E., Nabagło, I.: A modification of ELS algorithm for bilinear time-series model identification. ZeszytyNaukowePolitechnikiŚląskiej: Automatyka 108, 7–24 (1994)
9. Brunner, A., Hess, G.: Potential problems in estimating bilinear time-series models. Journal of Economic Dynamics and Control 19, 663–681 (1995)
10. Maliński, Ł., Bielińska, E.: Statistical Analysis of Minimum Prediction Error Variance in the Identification of a Simple Bilinear Time-Series Model. In: Advances in System Science, pp. 183–188. Academic Publishing House EXIT (2010)
11. Maliński, Ł.: On identification of coefficient of indivertible elementary bilinear time-series model. In: Proceedings XIV Symposium: Fundamental Problem of Power Electronics Electromechanics and Mechatronics PPEEm, pp. 194–196 (2011)
12. Maliński, Ł.: The Evaluation of Saturation Level for SMSE Cost Function in Identification of Elementary Bilinear Time-Series Model. In: 17 International Conference on Methods and Models in Automation and Robotics (2012)
13. Maliński, Ł., Figwer, J.: On stationarity of bilinear time-series. In: The 16th International Conference on Methods and Models in Automation and Robotics (2011)

Direction–of–Arrival Estimation in Nonuniform Noise Fields: A Frisch Scheme Approach

Roberto Diversi, Roberto Guidorzi, and Umberto Soverini

DEI – Department of Electrical, Electronic and Information Engineering
Viale del Risorgimento 2, 40136 Bologna, Italy
{roberto.diversi,roberto.guidorzi,umberto.soverini}@unibo.it

Abstract. This paper proposes a two-step identification procedure for the direction-of-arrival estimation problem in the presence of nonuniform white noise. The first step consists in estimating the unknown sensor noise variances by exploiting the properties of the Frisch scheme. Once that the noise covariance matrix has been identified, the angles of arrival are computed by using the classical ESPRIT algorithm. The effectiveness of the whole procedure is tested by means of Monte Carlo simulations.

Keywords: Direction–of–Arrival estimation, nonuniform noise, Frisch scheme.

1 Introduction

The estimation of the Directions-of-Arrival (DOAs) of multiple plane waves with narrow-band arrays of sensors is one of the central problems in radar, sonar, navigation, geophysics and acoustics applications. This problem has been studied intensively and several high resolution narrowband DOA estimators, such as MUSIC and the maximum likelihood (ML) method, have been proposed and analyzed in the past decades [1–6]. The maximum likelihood estimator shows excellent asymptotic performances and plays an important role in the context of these techniques [3, 6]. Many of the proposed ML estimators rely on the uniform white noise assumption, i.e. the sensor noises are assumed to be spatially uncorrelated white Gaussian noises with equal and unknown variance.

In many applications this assumption may be unrealistic and the sensor noise should be considered as a colored process, as discussed in [7, 8]. Nevertheless, in some real applications, for example when reverberating or seismic problems are modeled on the basis of measures obtained from sparse arrays, the general colored noise assumption can be simplified by assuming the sensor noises as spatially white. Reverberations, sensor imperfections, calibration errors and external noise suggest, however, considering unequal noise variances. In order to deal with the previous context the ML estimator has been recently rediscovered and both the deterministic and stochastic cases have been analyzed [8, 9]. The obtained results are quite accurate but the computational burden is non negligible.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_73, © Springer International Publishing Switzerland 2014

In this paper, the problem of DOA estimation in the presence of spatially non uniform independent sensor noise is solved by means of a two-step procedure. The first step consists in estimating the unknown sensor noise variances and exploits the properties of the Frisch scheme [10]. In the Frisch scheme context, the solution of the noise variance estimation problem is searched within a locus of solutions compatible with the covariance matrix of the noisy data by means of a suitable selection criterion. The proposed criterion is based on the rank deficiency property of a covariance matrix whose entries depend on the emitter signals covariances and the array transfer matrix. Once that the noise covariance matrix has been identified, the angles of arrival are computed by using the classical ESPRIT algorithm [4]. The effectiveness of the whole procedure is tested by means of Monte Carlo simulations.

The organization of the paper is the following. Section 2 defines the DOA estimation problem. Section 3 recalls some important properties of the Frisch scheme whereas the two-step DOA estimation procedure is described in Section 4. Section 5 shows the results of some Monte Carlo simulations while some concluding remarks are finally given in Section 6.

2 Signal Model and Problem Statement

Consider an array of n sensors receiving p narrow–band signals from sources with unknown DOAs θ_i $(i = 1, \ldots, p)$. The sources are assumed to be complanar and located in the far field of the array. The sensor array outputs are collected in a n–dimensional vector $y(t)$ and modeled by the following equation

$$y(t) = A(\theta)\, x(t) + e(t), \qquad t = 1, \ldots, N \tag{1}$$

where

$$\theta = [\theta_1, \theta_2, \ldots, \theta_p]^T, \tag{2}$$

N is the number of observations, $A(\theta)$ is the $n \times p$ array transfer matrix, $x(t)$ is the p–dimensional vector of source signals and $e(t)$ is the n–dimensional vector of the noises affecting the measures.
The additive noise $e(t)$ is assumed as a zero–mean spatially and temporally white Gaussian process with unknown diagonal covariance matrix

$$Q = E[e(t)e^H(t)] = \text{diag}\,[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]\ , \tag{3}$$

where $(\cdot)^H$ denotes Hermitian transpose and $E[\,\cdot\,]$ is the expectation operator.
The source signal $x(t)$ is a zero–mean, second–order ergodic random vector with non–singular $(p \times p)$ covariance matrix

$$R_x = E[x(t)x^H(t)]\ . \tag{4}$$

In the following, the number of sources p is assumed as known. The signal $x(t)$ is also assumed to be uncorrelated with the noise $e(t)$, so that the $(n \times n)$ array covariance matrix is given by

$$R_y = E[y(t)y^H(t)] = A(\theta)\, R_x\, A(\theta)^H + Q\ . \tag{5}$$

The problem under investigation is the following.

Problem 1. Given a set of N observations, collected in n–dimensional vectors $y(1), \ldots, y(N)$, estimate the noise variances σ_i^2 $(i = 1, \ldots, n)$ and the p angles of arrival θ_k $(k = 1, \ldots, p)$.

3 Some Properties of the Frisch Scheme

Let us introduce the following notation

$$\Sigma = R_y = E[y(t)\, y^H(t)] \tag{6}$$

$$\Sigma_0 = A(\theta)\, R_x\, A(\theta)^H \tag{7}$$

$$\tilde{\Sigma} = Q = E[e(t)\, e^H(t)] \ , \tag{8}$$

so that equation (5) can be rewritten as follows

$$\Sigma = \Sigma_0 + \tilde{\Sigma} \ . \tag{9}$$

Under the stated assumptions, this equation can be explained as follows: Σ is the $(n \times n)$ positive definite covariance matrix of the observation vector, $\tilde{\Sigma}$ is the diagonal covariance matrix of the noise, with unknown entries σ_i^2 $(i = 1, \ldots, n)$ and Σ_0 is the unknown covariance matrix of the noise-free data.
With reference to equation (9), the following mathematical problem has been deeply investigated in the literature with the name of Frisch scheme [10].

Problem 2. Given a $(n \times n)$ symmetric positive definite covariance matrix Σ, find all diagonal matrices $\tilde{\Sigma} = \mathrm{diag}\,[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$ with nonnegative elements σ_i^2 $(i = 1, \ldots, n)$ such that matrix $\Sigma - \tilde{\Sigma}$ is singular and nonnegative definite, i.e.

$$\Sigma_0 = \Sigma - \tilde{\Sigma} \geq 0 \quad \text{and} \quad \det \Sigma_0 = 0 \ . \tag{10}$$

Every positive definite or semidefinite diagonal matrix $\tilde{\Sigma}$ satisfying (10) is a *solution* of the Frisch Scheme. The corresponding point $P = (\sigma_1^2, \ldots, \sigma_n^2) \in \mathcal{R}^n$ can be considered as an admissible solution in the *noise space.* The locus of all the admissible solutions is described by the following theorem [11].

Theorem 1. All admissible solutions in the noise space lie on a convex (hyper)surface $\mathcal{S}(\Sigma)$ whose concavity faces the origin and whose intersections with the coordinate axes are the points $(0, \ldots, \sigma_i^2, \ldots, 0)$ corresponding to the n least squares solutions (see Fig. 1).

Definition 1. [12] The (hyper)surface $\mathcal{S}(\Sigma)$ will be called *singularity (hyper) surface* of Σ because its points define noise covariance matrices $\tilde{\Sigma}$ associated with singular matrices Σ_0.

Every point P of $\mathcal{S}(\Sigma)$ can thus be associated with a diagonal matrix $\tilde{\Sigma}(P)$ that, in turn, leads to a singular matrix

$$\Sigma_0(P) = \Sigma - \tilde{\Sigma}(P); \tag{11}$$

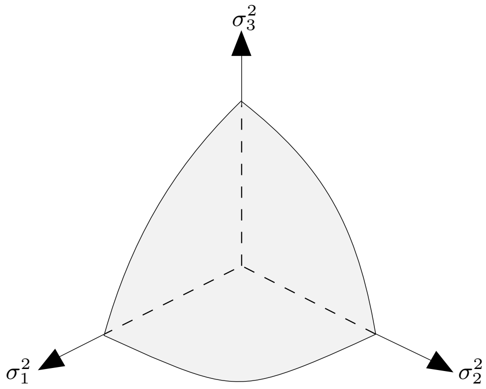


Fig. 1. Locus $\mathcal{S}(\Sigma)$ of admissible noise points for $n = 3$

moreover, rank $[\Sigma_0(P)] = m$ may change by varying P, i.e. $0 < m < n$ [13].
The corank of the singular matrix $\Sigma_0(P)$ is defined as $n - m$ and coincides with the dimension r of the subspace Ker $[\Sigma - \tilde{\Sigma}(P)]$. The maximum value that r can assume by varying $\tilde{\Sigma}(P)$, according with condition (10), is defined as the maximal corank of Σ

$$\text{Maxcor}_F(\Sigma) = \max_{\tilde{\Sigma}(P)} \left[n - \text{rank}(\Sigma - \tilde{\Sigma}(P))\right] \,, \qquad (12)$$

and represents the only invariant of the problem, that is the only feature that can be univocally identified from the noisy covariance matrix. Fundamental results concerning the evaluation of $\text{Maxcor}_F(\Sigma)$ are the following.

Theorem 2. [14] $\text{Maxcor}_F(\Sigma) = 1$ if and only if all entries of Σ^{-1} are positive or can be made positive (Frobenius–like according to the definition of Kalman) by changing the sign of some variables.

Theorem 3. [15] When $\text{Maxcor}_F(\Sigma) > 1$, $\mathcal{S}(\Sigma)$ is nonuniformly convex.

Theorem 4. [13] All points of $\mathcal{S}(\Sigma)$ where corank $(\Sigma) = k$ $(k > 1)$ are accumulation points for points where corank $(\Sigma) = k - 1$.

3.1 Radial Parametrization for Frisch Singularity Hypersurfaces [16]

A radial parametrization can be used effectively for computing the points of $\mathcal{S}(\Sigma)$ and also to perform fast searches on $\mathcal{S}(\Sigma)$ to minimize a given cost function. It is important to note that such a minimization can be performed by computing only the points requested by the adopted search procedure.
Let $\xi = (\xi_1, \ldots, \xi_n)$ be a generic point in the first orthant of $\mathcal{R}^n$ and ρ the straight line from the origin through ξ; the intersection, $P = (\sigma_1^2, \ldots, \sigma_n^2)$, between ρ and $\mathcal{S}(\Sigma)$ satisfies the conditions

$$\Sigma - \tilde{\Sigma}(P) \geq 0, \qquad \det\bigl(\Sigma - \tilde{\Sigma}(P)\bigr) = 0 \qquad (13)$$

and

$$\lambda\, P = \xi \quad \text{with} \quad \lambda > 0. \tag{14}$$

It follows that

$$\det\left(\Sigma - \frac{1}{\lambda}\, \tilde{\Sigma}^{\xi} \right) = 0 \tag{15}$$

where

$$\tilde{\Sigma}^{\xi} = \mathrm{diag}\left[\, \xi_1,\, \ldots, \xi_n \,\right]. \tag{16}$$

Relation (15) is equivalent ($\Sigma > 0$) to

$$\det\left(\lambda\, I - \Sigma^{-1} \tilde{\Sigma}^{\xi} \right) = 0 \tag{17}$$

so that the solution compatible with conditions (13) is given by

$$P = \frac{\xi}{\lambda_M} \tag{18}$$

with

$$\lambda_M = \max\, \mathrm{eig}\left(\Sigma^{-1} \tilde{\Sigma}^{\xi} \right). \tag{19}$$

The points of $\mathcal{S}(\Sigma)$ associated with straight lines from the origin can thus be obtained by computing Σ^{-1} and the intersection between any line and $\mathcal{S}(\Sigma)$ by means of (18) and (19).

4 Direction of Arrival Estimation

Given the covariance matrix Σ, the first step for the solution of Problem 1 consists in finding the covariance matrix $\tilde{\Sigma} = Q$ such that

$$\Sigma_0 = \Sigma - \tilde{\Sigma} \geq 0, \qquad \mathrm{rank}\big(\Sigma_0\big) \ = p\ . \tag{20}$$

In the Frisch scheme context this problem can be solved by searching for the points on the (hyper)surface $\mathcal{S}(\Sigma)$ with corank$\,(\Sigma) = n - p$, that is the points $P = (\sigma_1^2, \ldots, \sigma_n^2)$ on $\mathcal{S}(\Sigma)$ leading to matrices $\Sigma_0(P) = \Sigma - \tilde{\Sigma}$ having $n - p$ null eigenvalues. In the following, we will assume that

i) $n - p > 1$

ii) there is only one point P^* on $\mathcal{S}(\Sigma)$ such that $\Sigma_0(P^*) = \Sigma - \tilde{\Sigma}(P^*)$ has $n-p$ null eigenvalues.

For a discussion concerning assumption ii) see [17]. Note that, the point P^* is associated with the true noise variances so that

$$\Sigma_0(P^*) = \Sigma - \tilde{\Sigma}(P^*) = R_y - Q = A(\theta)\, R_x\, A(\theta)^H. \tag{21}$$

The above assumptions allow to introduce a selection criterion for finding the point P^* within the locus $\mathcal{S}(\Sigma)$. To this end, given a generic point $P \in \mathcal{S}(\Sigma)$

consider the singular value decomposition of the singular matrix $\Sigma_0(P) = \Sigma - \tilde{\Sigma}(P)$:

$$\Sigma_0(P) = U\,\Lambda\,U^H \tag{22}$$

where matrix $\Lambda = \operatorname{diag}[\lambda_1, \ldots, \lambda_{n-1}, \lambda_n]$ contains the eigenvalues, arranged in ascending order $\lambda_1 = 0$, $\lambda_{n-1} \leq \lambda_n$, and U has columns given by the corresponding normalized eigenvectors. The cost function

$$J(P) = \left\| \Sigma_0(P) \left[U_2 \cdots U_{n-p} \right] \right\|_2^2 , \tag{23}$$

satisfies the following properties

i) $J(P) \geq 0$

ii) $J(P^*) = 0$.

so that it can be minimized on $\mathcal{S}(\Sigma)$ in order to find P^*.

In practice, the covariance matrix Σ is replaced by the sample estimate

$$\hat{\Sigma} = \frac{1}{N} \sum_{t=1}^{N} y(t)\, y^H(t) \tag{24}$$

so that the minimum of $J(P)$ will be no longer zero. The search procedure on $\mathcal{S}(\hat{\Sigma})$ can be performed by means of the following algorithm.

Algorithm 1

1. Compute the sample estimates $\hat{\Sigma}$ given by (24).
2. Start from a generic line ρ belonging to the first orthant of $\mathcal{R}^n$.
3. Compute, by means of (18) and (19), the intersection $P = (\sigma_1^2, \ldots, \sigma_n^2)$ between ρ and $\mathcal{S}(\hat{\Sigma})$.
4. Construct the matrix $\tilde{\Sigma}(P) = \operatorname{diag}[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$ and compute the matrix
$$\Sigma_0(P) = \hat{\Sigma} - \tilde{\Sigma}(P)\ . \tag{25}$$
5. Compute the SVD (22) of matrix $\Sigma_0(P)$.
6. Compute the value of the cost function (23).
7. Move to a new direction corresponding to a decrease of $J(P)$.
8. Repeat steps 3–7 until the point $\hat{P} = (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2)$ associated with the minimum of $J(P)$ is found. The estimates of the noise variances are given by the coordinates of $\hat{P}$.
9. Save the SVD of the matrix $\Sigma_0(\hat{P})$
$$\Sigma_0(\hat{P}) = \hat{U}\,\hat{\Lambda}\,\hat{U}^H\ . \tag{26}$$

The second step for the solution of Problem 1 consists in the determination of the p array parameters θ_k $(k = 1, \ldots, p)$. This solution can be found by means of the following TLS–ESPRIT algorithm [4, 18].

Algorithm 2

1. Let B the matrix containing the last p columns of matrix $\hat{U}$ at point 9 of Algorithm 1. Partition matrix B as follows

$$B = \begin{bmatrix} B_u \\ x \dots x \end{bmatrix} = \begin{bmatrix} x \dots x \\ B_d \end{bmatrix} . \tag{27}$$

2. Compute the SVD of matrix $[B_u \; B_d]$

$$\begin{bmatrix} B_u^H \\ B_d^H \end{bmatrix} [B_u \; B_d] = V\, S\, V^H . \tag{28}$$

3. Partition matrix V in four $p \times p$ submatrices

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} . \tag{29}$$

4. Compute the eigenvalues $\lambda_1, \dots, \lambda_p$ of the matrix

$$\Psi_{TLS} = -V_{12} V_{22}^{-1} . \tag{30}$$

5. Find the estimates of the directions of arrival $\hat{\theta}_k$ $(k = 1, \dots, p)$ as

$$\hat{\theta}_k = \angle \lambda_k . \tag{31}$$

5 Numerical Results

The proposed approach has been tested by considering a uniform linear array with omnidirectional sensors and half-wavelength interelement spacing. We assume that there are $p = 2$ sources with equal power and $n = 5$ sensors. The direction of arrivals are

$$\theta_1 = 7^\circ \qquad \theta_2 = 13^\circ .$$

Under these assumptions the array transfer matrix is given by [18]

$$A = \begin{bmatrix} 1 & 1 \\ e^{i\pi \sin 7} & e^{i\pi \sin 13} \\ \vdots & \vdots \\ e^{i4\pi \sin 7} & e^{i4\pi \sin 13} \end{bmatrix} .$$

The two sources are mutually uncorrelated complex white gaussian processes with unit variance whereas the sensor noise $e(t)$ has the covariance matrix

$$Q = \mu \, \mathrm{diag}\, [\, 0.1 \; 0.3 \; 0.4 \; 0.2 \; 0.5 \,] ,$$

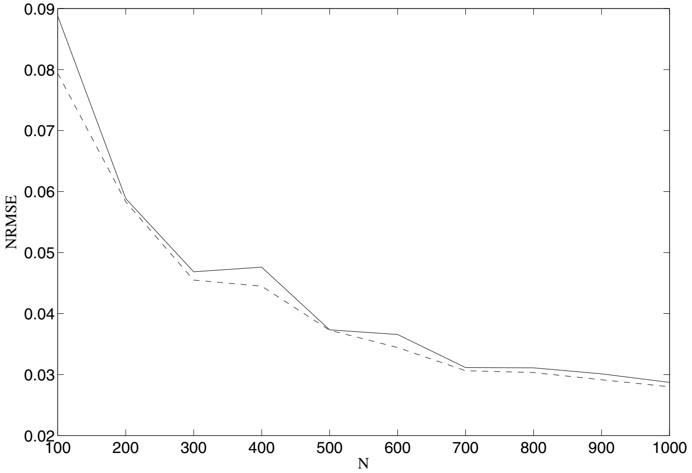


Fig. 2. NRMSE versus the number of samples (SNR = 10 dB): Frisch (solid line) and TLS (dashed line)

where the scalar $\mu > 0$ is adjusted in order to set the desired array signal to noise ratios defined as

$$\mathrm{SNR}_i = \frac{E[x_i(t)^2]}{n} \sum_{j=1}^{n} \frac{1}{\sigma_j^2}, \qquad i = 1, 2, \ldots, p.$$

Note that the worst noise power ratio $WNPR = \sigma^2_{max}/\sigma^2_{min}$ is 5 [8]. The proposed algorithm has been compared with the total least squares (TLS) approach that assumes the *a priori* knowledge of the noise covariance matrix up to a scalar [19]. The TLS solution is quite simple since the point $\hat{P}$ is directly computed by means of (18) and (19) using $\xi = [\,0.1\ 0.3\ 0.4\ 0.2\ 0.5\,]$.

In the first example the array SNR is fixed to 10 and the number of available samples is varied from $N = 100$ to $N = 1000$. For each value of N a Monte Carlo simulation of 100 independent runs has been performed. The algorithm performance has been evaluated by using the normalized root mean square error

$$\mathrm{NRMSE} = \frac{1}{\|\theta\|} \sqrt{\frac{1}{M} \sum_{i=1}^{M} \|\hat{\theta}^i - \theta\|^2}, \qquad (32)$$

where $\hat{\theta}^i$ denotes the estimate of the DOAs obtained in the i–th trial of the Monte Carlo simulation and M denotes the number of Monte Carlo runs. Figure 2 reports the NRMSE versus the number of data N for both Frisch and TLS. Table 1 reports the mean values and the associated standard deviations of the estimated angles of arrival (Frisch) for $N = 100, 500, 1000$ and SNR = 10 dB.

Table 1. True and estimated values of the directions of arrival (Frisch, SNR = 10 dB)

	true	$N = 100$	$N = 500$	$N = 1000$
θ_1	7°	7.2306 ± 0.8737	7.0154 ± 0.3638	6.9979 ± 0.2812
θ_2	13°	12.6687 ± 0.8985	12.9087 ± 0.4075	12.8704 ± 0.2925

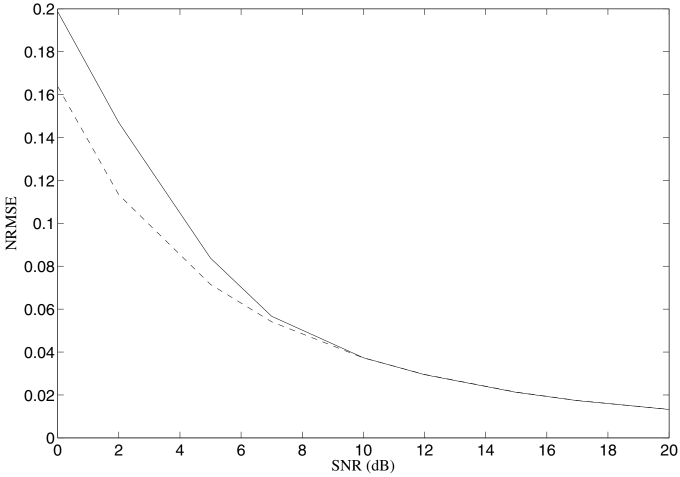


Fig. 3. NRMSE versus the array signal to noise ratio ($N = 500$): Frisch (solid line) and TLS (dashed line)

In the second example the number of samples is fixed to $N = 500$ whereas the array SNR ranges from 0 dB to 20 dB. For each value of the SNR a Monte Carlo simulation of 100 independent runs has been performed. Figure 3 reports the NRMSE versus the number of data N for both Frisch and TLS.

It can be noted that the proposed Frisch scheme approach leads to an estimation accuracy which is not so far from that of TLS without requiring the *a priori* knowledge on the noise variances.

6 Conclusion

In this paper, a new direction-of-arrival estimation approach has been proposed. The DOA estimation problem is solved by means of a two-step procedure where the first step is based on the properties of the Frisch scheme whereas the second one relies on the classical ESPRIT algorithm. The effectiveness of the method has been tested by means of Monte Carlo simulations and compared with that of the total least squares. The obtained results show that the new procedure leads to an estimation accuracy which is not so far from that of TLS without requiring any *a priori* knowledge on the noise variances.

References

1. Schmidt, R.O.: Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation 34, 276–280 (1986)
2. Bresler, Y., Macovski, A.: Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. IEEE Transactions on Acoustics, Speech and Signal Processing 34, 1081–1089 (1986)
3. Stoica, P., Nehorai, A.: MUSIC, maximum-likelihood and Cramer-Rao bound. IEEE Transactions on Acoustics, Speech and Signal Processing 37, 720–741 (1989)
4. Roy, R., Kailath, T.: ESPRIT–Estimation of signal parameters via rotational invariance techniques. IEEE Transactions on Acoustics, Speech and Signal Processing 37, 984–995 (1989)
5. Wax, M., Ziskind, I.: On unique localization of multiple sources by passive sensor arrays. IEEE Transactions on Acoustics, Speech and Signal Processing 37, 996–1000 (1989)
6. Stoica, P., Nehorai, A.: Performance study of conditional and unconditional direction-of-arrival estimation. IEEE Transactions on Acoustics, Speech and Signal Processing 38, 1783–1795 (1990)
7. Gershman, A.B., Matveyev, A.L., Böhme, J.F.: Maximum likelihood estimation of signal power in sensor array in the presence of unknown noise field. IEE Proceedings Radar, Sonar and Navigation 142, 218–224 (1995)
8. Pesavento, M., Gershman, A.B.: Maximum-likelihood direction-of-arrival estimation in the presence of unknown nonuniform noise. IEEE Transactions on Signal Processing 49, 1310–1324 (2001)
9. Chen, C.-E., Lorenzelli, F., Hudson, R.E., Yao, K.: Stochastic maximum-likelihood DOA estimation in the presence of unknown nonuniform noise. IEEE Transactions on Signal Processing 56, 3038–3044 (2008)
10. Guidorzi, R., Diversi, R., Soverini, U.: The Frisch scheme in algebraic and dynamic identification problems. Kybernetika 44, 585–616 (2008)
11. Beghelli, S., Guidorzi, R., Soverini, U.: The Frisch scheme in dynamic system identification. Automatica 26, 171–176 (1990)
12. Guidorzi, R.: Certain models from uncertain data: the algebraic case. Systems & Control Letters 17, 415–424 (1991)
13. Guidorzi, R.: Identification of the maximal number of linear relations from noisy data. Systems & Control Letters 24, 159–166 (1995)
14. Kalman, R.E.: Nine lectures on identification. LNEMS. Springer, Berlin
15. Schachermayer, W., Deistler, M.: The set of observationally equivalent errors–in–variables models. Systems & Control Letters 34, 101–104 (1998)
16. Guidorzi, R., Pierantoni, M.: A new parametrization of Frisch scheme solutions. In: Proc. of the XII Int. Conf. on Systems Science, Wroclaw, Poland, pp. 114–120 (1995)
17. Soverini, U., Beghelli, S.: Identification of static errors–in–variables models: the rank reducibility problem. Automatica 37, 1079–1084 (2001)
18. Stoica, P., Moses, R.: Introduction to Spectral Analysis. Prentice Hall, Upper Saddle River (1997)
19. Van Huffel, S., Vandewalle, J.: The Total Least Squares Problem: Computational Aspects and Analysis. SIAM, Philadelphia (1991)

Closed-Loop System Identification Using Quantized Observations and Integrated Bispectra

Teresa Główka and Jarosław Figwer

The Silesian University of Technology, 44-100 Gliwice, Akademicka 16, Poland
{Teresa.Glowka,Jaroslaw.Figwer}@polsl.pl

Abstract. The aim of this paper is to present a novel approach to closed-loop discrete-time system frequency response and the corresponding parametric model identification using repeated discrete-time non-Gaussian excitation and quantized plant output observartions. The specially designed identification experiment based on data averaging is proposed to reduce the quantization effect and enhance signal-to-noise ratio. The integrated bispectra-based identification method is proposed to handle with closed-loop system identification problems. A focus on model identification in the case of disturbance-free plant output and output signal level comparable with data acquisition system accuracy is given. Convergence of the identified model to true plant is discussed. The discussion is illustrated by an example showing properties of the presented approach.

Keywords: system identification, closed-loop identification, identification experiment, higher order spectra.

1 Introduction

System identification based on processing of sampled signals is widely discussed in the literature (see e.g. [1–3]). The problem of plant identification in closed-loop systems – in general much more complicated than the open-loop identification – is also presented in many publications (e.g. [4–11]). Unfortunately, the fact that the sampled signals are obtained from A/D converters which inherent parts are quantizers, is usually ignored. Because these quantizers can have a great influence on identification results, there is a need to adapt the system identification techniques to the case of use quantized observations. Recently, some modifications of the system identification techniques dealing with quantized observations processing have been proposed, e.g. [12–20]. However, the problem of closed-loop system identification from quantized observations is not considered therein, except [13] and [21].

The purpose of this paper is to discuss new ideas of discrete-time frequency response as well as the corresponding parametric model identification for linear discrete-time dynamic single input single output (SISO) plants operating in closed-loop systems, using deterministic (in sense of periodicity) discrete-time excitations being realisations of non-Gaussian random processes and quantized

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_74,

observations. To show an influence of quantization, the discussion is concentrated on model identification in the case of disturbance-free plant output. Moreover, the output signal level is assumed to be comparable with data acquisition system accuracy. A focus on convergence of the identified model to true plant is given.

The paper is organized as follows: in Section 2 the linear discrete-time dynamic SISO plant identification problem is stated, the plant works in feedback control system and quantized observations are taken for identification; in Section 3 data acquisition and initial data processing issues are presented, ideas of frequency response and parametric model identification are reminded, and a focus on quantized observations processing is given; in Section 4 the presented discussion is illustrated by some examples.

2 Problem Formulation

In the presented discussion a closed-loop linear time-invariant discrete-time dynamic SISO system is considered, see block diagram in Fig. 1. The following notations of signals and transfer functions are used in Fig. 1 and in the sequel: the plant $H(z^{-1})$ is the unknown rational transfer function to be identified, $C(z^{-1})$ is the controller, $u(i)$ is the known input signal, $y(i)$ is the measurable output signal, $d(i)$ represents disturbances influencing the plant, $w(i)$ is the set-point value (assumed to be zero), $v(i)$ is the external excitation signal introduced to the system for the purpose of identification.

For control systems, the structure presented in Fig. 1 has fundamental importance. Identification in the closed-loop system is sometimes falsely taken for errors-in-variables problem. However, this closed-loop system structure is significantly different in nature from the typical errors-in-variables system, because the disturbance at the plant output is transmitted as well to the plant input through feedback loop. It causes that the disturbances influencing both input and output signals are correlated, and this fact complicates the identification task.

The theoretical analysis given in [10, 22] leads to the conclusion, that under some conditions (listed in the sequel) the frequency response identification methods based on higher-order spectra may be successfully applied in the case of correlated disturbances, because the influence of disturbances is theoretically nullified in the higher-order spectra domain. Here, by higher-order spectra,

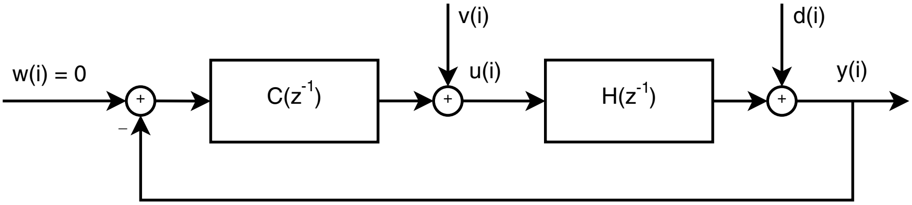


Fig. 1. Closed-loop system block diagram

frequency-domain representations of statistics named cumulants are understood. It is widely known [23, 24], that all cumulants of orders higher than second are identically equal zero for Gaussian random processes. Moreover, for odd-order cumulants the identity to zero concerns not only Gaussian processes, but also all non-skewed processes (i.e. for random processes consisting of random variables with probability density functions symmetric around mean value). This property allows for theoretical elimination of additive Gaussian noise.

The obvious cost of higher-order spectra-based identification methods is that the higher is order of used spectra, the higher is also the variance of obtained estimates. In the case of low signal-to-noise ratio the variance can be so large, that the estimates quality will be very poor (even worse than for the estimates obtained by second-order spectra based method, i.e. classical spectral analysis). Therefore, in general, data sequences for higher-order spectra-based methods should be longer than for classical methods. The second drawback is that use of higher-order spectra requires greater computational effort than classical spectral analysis. It can be overcome by use of integrated higher-order spectra instead of ordinary non-integrated ones, because the accuracy of both estimates is similar, but the computational load is dramatically lower for integrated spectra-based methods and it is comparable to the classical spectral analysis [10, 22].

In this paper, the influence of quantizers on identification results is invoked. The quantizers are internal parts of A/D converters and an error produced by them can be interpreted as an additive random noise uniformly distributed over the range $[-q/2, +q/2]$, where q denotes a single quant value. Hence, if the identification method is based on odd-order cumulant spectra, the quantization influence on identification results is supposed to be substantially reduced.

The method proposed in the paper is based on integrated bispectra (i.e., integrated third-order spectra). The general assumptions for this method are:

- the entire system (together with controller part) is causal and asymptotically stable;
- the disturbances are stationary random processes with third-order spectra equal to zero;
- the external excitation signal $v(i)$ is a realization of a stationary and ergodic random process with third-order spectrum different from zero, and independent of disturbances.

Additionally, it is assumed that:

- the input signal (together with the excitation) is generated using a D/A converter equipped with zero-order hold filter in such a way that quantization effects of the excitation generation system can be neglected;
- in the data acquisition system there is an A/D converter equipped with an uniform quantizer – it implies that nonlinearity of the data acquisition system must be taken into account [13, 14].

Hence, instead of non-quantized $y(i)$, the quantized output signal $y_{\mathrm{q}}(i)$ is taken for identification. This quantized output $y_{\mathrm{q}}(i)$ is turned back to the input through the feedback loop, so the input signal $u(i)$ also contains the error coming from quantization of output. Therefore it is denoted as $u_{\mathrm{q}}(i)$ in the sequel.

The goal of identification is to calculate estimate of unknown transfer function $H(z^{-1})$, having finite length sequences of input $u_{\mathrm{q}}(i)$ and output $y_{\mathrm{q}}(i)$ signals. In the proposed integrated bispectra-based method, the obtained model is in the form of a set of system frequency response values $\hat{H}(j\Omega n)$ computed for a fixed set of discrete relative frequencies Ωn in the range $[0, \pi]$. The rational transfer function $\hat{H}(z^{-1})$ parameters can be calculated from the obtained frequency response estimates $\hat{H}(j\Omega n)$ as well. The data are acquired during identification experiments in which an additional discrete-time excitation $v(i)$ is used. The external excitation is a single realization of non-Gaussian non-symmetrically distributed (skewed) random process.

3 Identification Procedure

3.1 Data Acquisition and Preprocessing

During identification experiment [4, 13, 15] start of data acquisition is delayed with respect to the instant of putting the excitation at the linear discrete-time dynamic SISO system input. It starts after all transients have decayed. Under such discrete-time dynamic SISO system steady-state conditions the N-sample discrete-time external excitation $v(i)$ is repeated m times and added to the controller output. To estimate models, the linear discrete-time dynamic SISO system input $u_{\mathrm{q}}(i)$ and output $y_{\mathrm{q}}(i)$ signals are represented by the following set of values: $\{u_{\mathrm{q}}(0), u_{\mathrm{q}}(1), \ldots, u_{\mathrm{q}}(mN-1)\}$ and $\{y_{\mathrm{q}}(0), y_{\mathrm{q}}(1), \ldots, y_{\mathrm{q}}(mN-1)\}$.

It is assumed in the presented discussion, that prior to quantization independent realisations of the random variable, uniformly distributed in the range covering the quant of A/D converter, are added to processed values. This randomized quantization [26] gives an opportunity to reduce an influence of quantization effects on identification results. It is of particular importance in the case of model identification for disturbance-free plants excited by deterministic (in sense of periodicity) discrete-time excitations and signal levels comparable with data acqusition system accuracy.

To identify models of the linear discrete-time dynamic SISO system, the data sets are initially processed. This preprocessing transforms the mN-sample data sequences into the following N-sample sets: $\{\tilde{u}_{\mathrm{q}}(0), \tilde{u}_{\mathrm{q}}(1), \ldots, \tilde{u}_{\mathrm{q}}(N-1)\}$ and $\{\tilde{y}_{\mathrm{q}}(0), \tilde{y}_{\mathrm{q}}(1), \ldots, \tilde{y}_{\mathrm{q}}(N-1)\}$ according to the equations

$$\tilde{u}_{\mathrm{q}}(i) = \frac{1}{m}\sum_{l=0}^{m-1} u_{\mathrm{q}}(i+lN)\,, \qquad \tilde{y}_{\mathrm{q}}(i) = \frac{1}{m}\sum_{l=0}^{m-1} y_{\mathrm{q}}(i+lN)\,, \tag{1}$$

for each $i = 0, 1, \ldots, N-1$.

The obtained N-sample data sequences are unbiased and consistent estimators of the corresponding quantization-free discrete-time input and output signal values. Hence, their variances decline with the increase of the number m of processed N-sample data segments. Additionally, it is worth to emphasize that:

$$\lim_{m\to\infty} \tilde{u}_{\mathrm{q}}(i) = u(i)\,, \qquad \lim_{m\to\infty} \tilde{y}_{\mathrm{q}}(i) = y(i) \qquad a.s. \tag{2}$$

The above remarks imply that the discussed data acquisition and preprocessing is a tool that allows to reduce influence of quantization effects on identification results. It is also worth to emphasize that the randomized quantization is necessary to obtain the mentioned reduction. Without this quantization (i.e. using deterministic quantization) the initial data processing gives for all m values $\tilde{u}_{\mathrm{q}}(i) = u_{\mathrm{q}}(i)$ and $\tilde{y}_{\mathrm{q}}(i) = y_{\mathrm{q}}(i)$ $(i = 0, 1, \ldots, N-1)$.

3.2 Frequency Response Identification

Let $\tilde{U}_{\mathrm{q}}(j\Omega n)$ and $\tilde{Y}_{\mathrm{q}}(j\Omega n)$ denote N-point discrete Fourier transforms of the considered SISO system quantized and preprocessed input and output data sets, respectively:

$$\tilde{U}_{\mathrm{q}}(j\Omega n) = \sum_{i=0}^{N-1} \tilde{u}_{\mathrm{q}}(i) e^{-j\Omega n i}\,, \qquad \tilde{Y}_{\mathrm{q}}(j\Omega n) = \sum_{i=0}^{N-1} \tilde{y}_{\mathrm{q}}(i) e^{-j\Omega n i}\,, \tag{3}$$

where n denotes consecutive harmonics of the fundamental relative frequency $\Omega = 2\pi/N$.

The frequency response $\hat{H}(j\Omega n)$ $(n = 0, 1, \ldots, N/2)$ of the linear discrete-time dynamic SISO system can be estimated with the help of integrated bispectra-based method as [10, 22]

$$\hat{H}(j\Omega n) = \frac{\hat{IB}_{\tilde{u}_{\mathrm{q}}\tilde{y}_{\mathrm{q}}}(j\Omega n)}{\hat{IB}_{\tilde{u}_{\mathrm{q}}\tilde{u}_{\mathrm{q}}}(j\Omega n)}\,, \tag{4}$$

where the integrated bispectrum (denominator) and the integrated cross-bispectrum (numerator) estimators are given by

$$\hat{IB}_{\tilde{u}_{\mathrm{q}}\tilde{u}_{\mathrm{q}}}(j\Omega n) = \frac{1}{N} R^{*}_{2\tilde{u}_{\mathrm{q}}}(j\Omega n) \tilde{U}_{\mathrm{q}}(j\Omega n)\,, \tag{5}$$

$$\hat{IB}_{\tilde{u}_{\mathrm{q}}\tilde{y}_{\mathrm{q}}}(j\Omega n) = \frac{1}{N} R^{*}_{2\tilde{u}_{\mathrm{q}}}(j\Omega n) \tilde{Y}_{\mathrm{q}}(j\Omega n)\,, \tag{6}$$

where * denotes complex conjugation, and

$$R_{2\tilde{u}_{\mathrm{q}}}(j\Omega n) = \sum_{i=0}^{N-1} r_{2\tilde{u}_{\mathrm{q}}}(i) e^{-j\Omega n i} \tag{7}$$

is the N-point discrete Fourier transform of

$$r_{2\tilde{u}_{\mathrm{q}}}(i) = \tilde{u}^2_{\mathrm{q}}(i) - \frac{1}{N} \sum_{\nu=0}^{N-1} \tilde{u}^2_{\mathrm{q}}(\nu)\,. \tag{8}$$

The presented identification method exploits direct estimators of such integrated bispectra, i.e., it is based straightforwardly on discrete Fourier transforms of collected data sets. In this approach, each estimate of integrated bispectrum must be smoothed in the frequency domain, e.g. over neighbouring frequencies using rectangular frequency window with span $(2L+1)$. Otherwise, the empirical transfer function estimator is obtained [1, 22] and consequently the identification method wastes higher-order spectra-based profits. The estimators (5) and (6) are asymptotically unbiased but inconsistent. Smoothing in the frequency domain makes them consistent [10, 22, 24]. Hence, the frequency response estimator (4) is also consistent. This is true for closed-loop systems, too [10, 22].

3.3 Parametric Model Estimation

The frequency response estimate calculated from Eq. (4) is a set of complex values obtained for a set of frequencies. However, in many practical applications, a parametric model in the form of rational transfer function $\hat{H}(z^{-1}) = \hat{B}(z^{-1})/\hat{A}(z^{-1})$ is needed, where $\hat{B}(z^{-1})$ and $\hat{A}(z^{-1})$ are finite-order polynomials of z^{-1}. It can be achieved by the least-squares approximation (i.e, equation error formulation, see [25]) of the frequency response estimate. In this approach, the rational transfer function parameters are obtained by minimizing the square norm of the difference between frequency response of the searched parametric model and the estimated frequency response (nonparametric).

4 Exemplary Identification Results

The presented approach to identification of linear discrete-time dynamic SISO plants operating in closed-loop systems using the integrated bispectra-based method and the quantized observations is illustrated by identification results for the following unstable plant:

$$H(z^{-1}) = \frac{0.010z^{-1} + 0.005z^{-2}}{1.000 - 1.850z^{-1} + 0.525z^{-2}}, \tag{9}$$

stabilized by the controller $C(z^{-1}) = 35$. The excitation signal $v(i)$ was a N-sample single realisation of a discrete-time exponential random process with zero mean and variance equal to 0.1. During simulations, there was no disturbance $d(i)$ at the plant output. Instead of this, the plant output was assumed to be measured using A/D unit utilizing randomized quantization, with 8-bit converter which range was from 0 to 10 V. Therefore the plant output was quantized, and the quantization error was treated as a "disturbance" in the system. The chosen parameters of system and excitation caused that during identification experiments no more than 3 bits of A/D converter were used.

The proposed integrated bispecta-based method (IB) was compared with three other methods: classical spectral analysis (SA), least-square method (LS), and prediction error method (PE). In the literature, the last method is suggested as the best for closed-loop system identification [3]. For IB and SA, the least-squares approximation in the frequency-domain was used to estimate parameters of the rational transfer function model $\hat{H}(z^{-1})$.

Results of plant parameters identification for all four methods and different values of N and m are summarized in Table 1. Additionally the frequency responses of parametric models obtained for first three methods are presented in Fig. 2 for $m = 100$, $N = 8192$ (left plots) and $m = 10$, $N = 32768$ (right plots). The convergence of results given by proposed method is illustrated in Fig. 3 for fixed m and varying N (left plots) and for fixed N and varying m (right plots). Analysis of these results shows that:

- the results of all methods improve with increasing m, what is rather obvious, because m enhances signal-to-noise ratio;
- the results of IB method improve with increasing N, because this estimator is consistent; for SA and LS methods it is not observed (thought variance of estimates decreases with increase of N, they still produce bias);
- the use of proposed IB method instead of classical methods (SA and LS) brings profits especially when values of m are small, it is because the quantization error is than more significant;
- PE method works well in all cases; however, this method is time-consuming (identification takes several hundreds times longer than identification with other methods), moreover, this method is sensitive to model structure choice [22].

Table 1. Values of identified parameters for different identification methods

Method	Parameters				Method	Parameters			
	-1.8500	0.5250	0.0100	0.0050		-1.8500	0.5250	0.0100	0.0050
	$N = 1024$, $m = 10$					$N = 8192$, $m = 10$			
IB	-1.3373	0.0865	0.0083	0.0068	IB	-1.5043	0.2274	0.0093	0.0060
SA	-2.5147	0.7621	0.0282	-0.0002	SA	-2.6704	0.7816	0.0314	0.0005
LS	-1.1672	-0.2677	0.0099	0.0109	LS	-1.2091	-0.2371	0.0100	0.0110
PE	-1.8386	0.5205	0.0097	0.0051	PE	-1.8356	0.4827	0.0106	0.0052
	$N = 1024$, $m = 100$					$N = 8192$, $m = 100$			
IB	-1.8461	0.5329	0.0105	0.0046	IB	-1.8060	0.5145	0.0093	0.0048
SA	-1.8984	0.4993	0.0121	0.0051	SA	-1.9272	0.5387	0.0121	0.0048
LS	-1.6477	0.2750	0.0098	0.0070	LS	-1.6610	0.2872	0.0100	0.0069
PE	-1.8729	0.5580	0.0101	0.0045	PE	-1.8515	0.5270	0.0100	0.0050
	$N = 1024$, $m = 1000$					$N = 8192$, $m = 1000$			
IB	-1.8600	0.5259	0.0100	0.0051	IB	-1.8392	0.5114	0.0100	0.0051
SA	-1.8743	0.5324	0.0104	0.0048	SA	-1.8644	0.5311	0.0103	0.0049
LS	-1.8339	0.5080	0.0100	0.0050	LS	-1.8300	0.5000	0.0100	0.0051
PE	-1.8489	0.5227	0.0100	0.0050	PE	-1.8506	0.5256	0.0100	0.0050

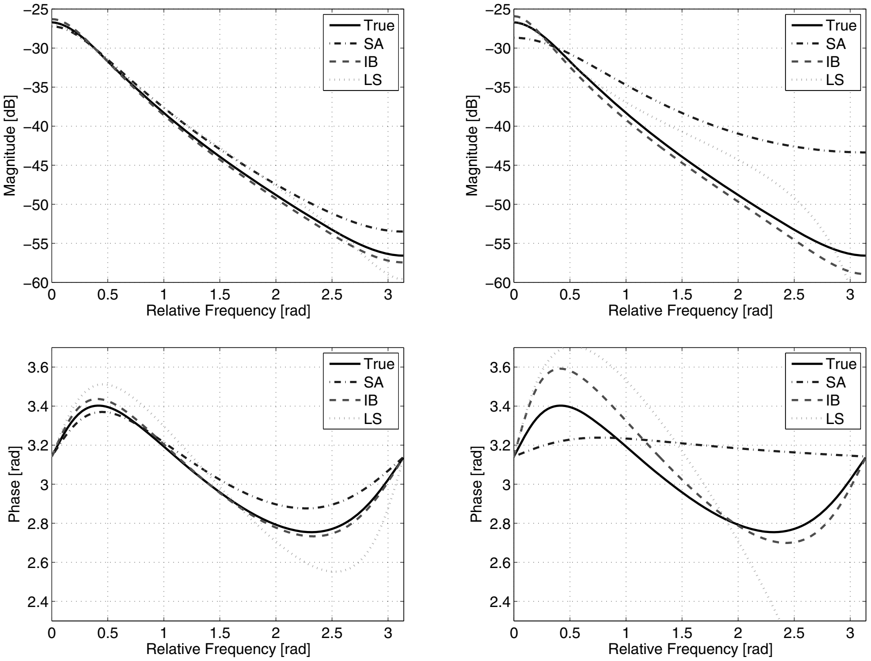


Fig. 2. Comparison of identification results (magnitudes and phases of frequency responses) for different identification methods; $m = 100$, $N = 8192$ (left) and $m = 10$, $N = 32768$ (right)

Concluding, it follows from the presented identification results that the discussed approach to model identification from quantized observations is a powerful tool that allows to estimate models when information about plant is hidden in between single multiplicities of A/D converter quant.

5 Summary

In the paper the approach to the closed-loop discrete-time SISO sytem frequency response and the corresponding parametric model identification using the repeated discrete-time non-Gaussian excitation and the quantized plant output observations is discussed. The specially designed identification experiment based on data averaging is proposed to reduce quantization effects and enhance signal-to-noise ratio. The integrated bispectra-based identification method is proposed to handle with closed-loop system identification problems.

The discussion is concentrated on the case of disturbance-free plant output, to extract the influence of quantization. However, the proposed method can be successfully applied for non-zero disturbances at the plant output.

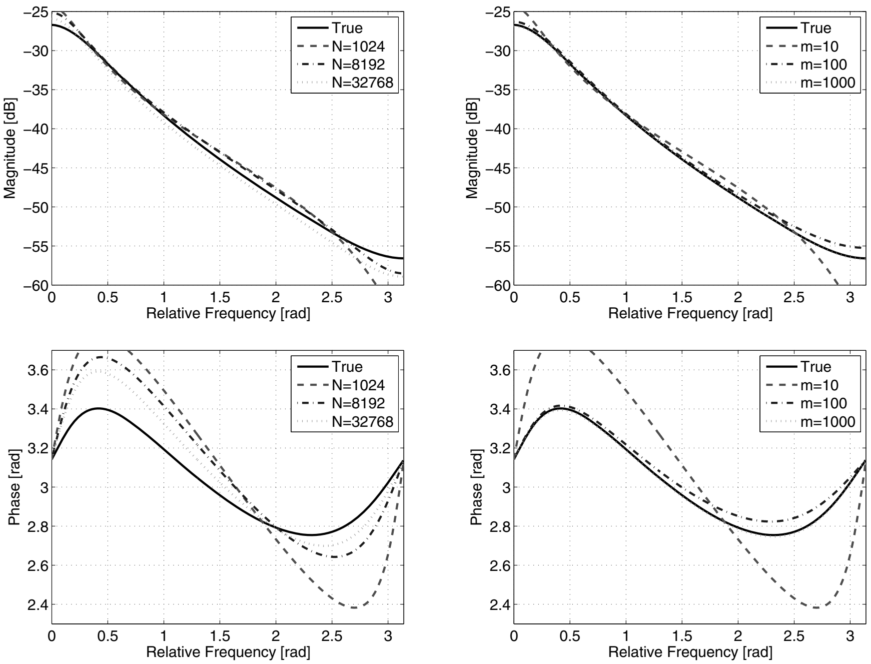


Fig. 3. Comparison of identification results (magnitudes and phases of frequency responses) for integrated bispectra-based method and different values of N or m; $m = 10$, different N (left) and $N = 1024$, different m (right)

Finally, it is worth to mention, that the presented approach will be valid for other system structures, in which disturbances influencing the input and output are correlated, e.g. for feedforward systems with disturbance compensation [22].

Acknowledgements. The partial financial support of this research by The Polish Ministry of Science and Higher Education is gratefully acknowledged.

References

1. Ljung, L.: System Identification – Theory for the User. Prentice Hall PTR, New Jersey (1999)
2. Pintelon, R., Schoukens, J.: System Identification. A Frequency Domain Approach. IEEE Press, New York (2001)
3. Söderström, T., Stoica, P.: System Identification. Prentice Hall, Ltd., UK (1994)
4. Figwer, J.: Closed-Loop System Identification with Multisine Excitation. In: Proceedings of the Eighth IEEE International Conference on Methods and Models in Automation and Robotics, Szczecin, Poland, pp. 477–482 (2002)
5. Gevers, M., Bombois, X., Hildebrand, R., Solari, G.: Optimal Experiment Design for Open and Closed-Loop System Identification. Communications in Information and Systems 11, 197–224 (2011)

6. Gustavsson, I., Ljung, L., Soderstrom, T.: Identification of Processes in Closed Loop – Identifability and Accuracy Aspects. Automatica 13, 59–75 (1977)
7. Hof, P.: Closed-Loop Issues in System Identification. Annual Reviews in Control 22, 173–186 (1998)
8. Ljung, L.: Identification in Closed Loop: Some Aspects on Direct and Indirect Approaches. In: Proceedings of the IFAC Symposium on System Identification, Fukuoka, Japan, pp. 141–146 (1997)
9. Forssell, U., Ljung, L.: Closed-loop identification revisited. Automatica 35, 1215–1241 (1999)
10. Tugnait, J.K., Zhou, Y.: On Closed-Loop System Identification Using Polyspectral Analysis Given Noisy Input-Output Time-Domain Data. Automatica 36, 1795–1808 (2000)
11. Zheng, X.W., Feng, C.B.: A Bias Correction for Indirect Identification of Closed-Loop Systems. Automatica 31, 1019–1024 (1995)
12. Aguero, J.C., Goodwin, G.C., Yuz, J.I.: System identification using quantized data. In: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, pp. 4263–4268 (2007)
13. Figwer, J.: Continuous-Time Dynamic System Identification with Multisine Random Excitation Revisited. Archives of Control Sciences 20, 133–149 (2010)
14. Figwer, J.: Frequency Response Identification in the Case of Periodic Disturbances. In: The 16th International Conference on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland, pp. 1–4 (2011)
15. Figwer, J.: Model Identification Using Quantized Data – a Distrubance Free Case. In: The 17th International Conference on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland, pp. 240–243 (2012)
16. Godoy, B.I., Goodwin, G.C., Aguero, J.C., Marelli, D., Wigren, T.: On identification of FIR systems having quantized output data. Automatica 47, 1905–1915 (2011)
17. Gustafsson, F., Karlson, R.: Statistical results for system identification based on quantized observations. Automatica 45, 1794–2801 (2009)
18. Suzuki, H., Sugie, T.: System identification based on quantized I/O data corrupted with noises. In: Proceedings of the 17th International Symposium on Mathematical Theory and Systems, Kyoto, Japan (2006)
19. Wang, L.Y., Yin, G.G., Zhang, J.F., Zhao, Y.: System Identification with Quantized Observations. Birkhauser, Boston (2010)
20. Widrow, B., Kollar, I.: Quantization Noise: Roundoff error in digital computation, signal processing, control, and communications. Cambridge University Press (2008)
21. Wang, M., Thornhill, N.F., Huang, B.: Closed Loop Identification Based on Quantization. In: Proceedings of the 15th IFAC World Congress, Barcelona, paper 1331 (2002)
22. Główka, T.: Higher Order Spectra for Frequency Model Identification. Jacek Skalmierski Computer Studio, Gliwice (2011)
23. Nikias, C.L., Mendel, J.M.: Signal Processing with Higher-Order Spectra. IEEE Signal Processing Magazine 10(4), 1–15 (1993)
24. Nikias, C.L., Petropulu, A.P.: Higher-Order Spectra Analysis – A Nonlinear Signal Processing Framework. PTR Prentice Hall Inc., Englewood Cliffs (1993)
25. Tugnait, J.K., Ye, Y.: Stochastic system identification with noisy input-output measurements using polyspectra. IEEE Transactions on Automatic Control AC-40, 670–683 (1995)
26. Bilinskis, I.: Digital Alias-Free Signal Processing. John Wiley & Sons (2007)

Identification of Fractional-Order Continuous-Time Hybrid Box-Jenkins Models Using Refined Instrumental Variable Continuous-Time Fractional-Order Method

Walid Allafi and Keith J. Burnham

Control Theory and Applications Centre, Coventry University, Coventry CV1 5FB, UK
{allafiw,ctac}@coventry.ac.uk

Abstract. This paper illustrates the identification of a Box-Jenkins model from sampled input and output data. This is achieved by extending a refined instrumental variable continuous-time (RIVC) method to a refined instrumental variable continuous-time fractional-order (RIVCF) method. The model is a hybrid of continuous and discrete-time as well as fractional and integer-orders. The model consists of a fractional-order linear continuous-time (FLC) transfer function and noise. The FLC transfer function represents the noise free system and the noise represents an integer-order discrete-time autoregressive moving average (ARMA). Monte Carlo simulation analysis is applied for illustrating the performance of the proposed RIVCF method.

Keywords: Fractional order, System identification, Box-Jenkins models, Refined instrumental variables.

1 Introduction

Although fractional calculus was defined approximately three hundred years ago by Riemann and Liouville, it has only been realised in real-world applications in the last two decades. This has been due to a massive increase in computer technology which eases the numerical simulation of fractional systems. Some physical systems are modelled as a fractional-order system such as the diffusion process in a battery cell [9] and heat transfer systems [2].

There have been developments in system identification and parameter estimation by applying instrumental variable (IV) approaches since the 1960s as illustrated in [5]. A continuous-time (CT) system may be directly identified based on sampled input and output data [4]. The refined instrumental variable continuous-time (RIVC) method was proposed by P.C. Young for identifying a hybrid Box-Jenkins model from sampled input and output [6]. The model is a hybrid of a continuous-time system and discrete-time noise process.

The objective of this paper is to extend the RIVC method to refined instrumental variable continuous-time fractional-order (RIVCF) for identifying a hybrid Box-Jenkins model from sampled input and output data. The model is a hybrid of a fractional-order continuous-time system and discrete-time noise process.

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_75, © Springer International Publishing Switzerland 2014

2 Fractional-Order Model

A fractional differential equation can describe the fractional-order model in the following manner:

$$\begin{aligned} &a_0 D^{\alpha_n} x(t) + a_1 D^{\alpha_{n-1}} x(t) + \ldots + a_n D^{\alpha_0} x(t) \\ &= b_0 D^{\beta_m} u(t) + b_2 D^{\beta_{m-1}} u(t) + \ldots + b_m D^{\beta_0} u(t) \end{aligned} \tag{1}$$

where $x(t)$ and $u(t)$ are the output and input of the model, respectively. $D^{\alpha}x(t) = \frac{d^{\alpha}x(t)}{dt^{\alpha}}$,$a_j(j = 0,1,\ldots n)$, $b_j(j = 1,2,\ldots m)$ are constants,$\alpha_j(j = n, n-1,\ldots 0)$, $\beta_j(j = m, m-1,\ldots 0) \in \mathbb{R}^+and\alpha_n > \alpha_{n-1} \ldots > \alpha_0$and $\beta_m > \beta_{m-1} \ldots > \beta_0$.

The fractional derivative of order $\alpha \in \mathbb{R}^+$ was defined by Riemann-Liouville [3] as:

$$D^{\alpha}x(t) \triangleq D^m I_0^{m-\alpha} x(t) = \frac{d^m}{dt^m}\left[\frac{1}{\Gamma(m-\alpha)}\int_0^t (t-\tau)^{-(\alpha-m+1)} x(\tau)\,d\tau\right] \tag{2}$$

where $m-1 < \alpha < m, m \in \mathbb{Z}$, and the Euler function is defined as:

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt \qquad \forall x \in \mathbb{R}$$

A discrete-time definition of the concept of fractional differentiation was defined by Grünwald–Letnikov based on the generalisation of the backward difference [1]:

$$D^{\alpha}x(t)\Big|_{t=kh} = \lim_{h\to 0}\frac{1}{h^{\alpha}}\sum_{j=0}^{k}(-1)^j\binom{\alpha}{j}x(kh-jh) \tag{3}$$

where $\binom{\alpha}{j}$ is Newton's binomial function. It is generalised using the Euler function, and extended to fractional-order:

$$\binom{\alpha}{j} = \frac{\Gamma(\alpha+1)}{\Gamma(j+1)\Gamma(\alpha-j+1)}$$

2.1 Numerical Simulation

In order to obtain a numerical approximation of factional derivatives, the Grünwald–Letnikov definition (3) is used:

$$D^{\alpha}x(t) \approx \Delta_h^{\alpha} x(t)$$

$$\Delta_h^{\alpha} x(t)\Big|_{t=kh} = \frac{1}{h^{\alpha}} \sum_{j=0}^{k} \omega_j^{\alpha} x(kh - jh) \tag{4}$$

where

$$\omega_j^{\alpha} = (-1)^j \binom{\alpha}{j}$$

It can be noted from (4) that as time, denoted t, increases, there is a need to increasingly add more and more to the summation for computing the solution. However it is observed that for large t the coefficients of the Grünwald–Letnikov definition for the more recent values have much larger influence than the older values. Hence the numerical solution can be approximated by using the only recent values. This leads to a memory length, denoted L, [1], such that:

$$D^{\alpha} x(t) \approx_{t-L} \Delta^{\alpha} x(t), t > L$$

2.2 Fractional-Order Transfer Function and State Space Representation

The Laplace transform of fractional-order derivative [3]:

$$\mathcal{L}\left(D^{\alpha} f(t)\right) = s^{\alpha} F(s) \text{ if } f(t) = 0 \ \forall t < 0$$

Applying this property on (1) yields the fractional-order transfer function (FOTF):

$$G(s) = \frac{X(s)}{U(s)} = \frac{b_0 s^{\beta_m} + \ldots + b_m s^{\beta_0}}{a_0 s^{\alpha_n} Y(s) + \ldots + a_n s^{\alpha_0}} \tag{5}$$

where $x(t)$ and $u(t)$ are relaxed at $t = 0$.

The FOTF is termed commensurate iff the orders of the derivative in (1) are integer multipliers of a base order, denoted α. Therefore (5) becomes:

$$G(s) = \frac{\sum_{k=0}^{m} b_{m-k} s^{k\alpha}}{\sum_{k=0}^{n} a_{n-k} s^{k\alpha}}$$

A fractional-order state space (FOSS) representation may be obtained by converting the FOTF. It is possible only when the FOTF is commensurate [1]. The obtained FOSS representation becomes [1]:

$$s^{n} X(s) = AX(s) + BU(s)$$

$$Y(s) = CX(s) + DU(s)$$

where $X(s)$ is the state variable vector and A, B, C and D are system, input, output and feed through matrices, respectively.

Converting a FOSS representation to a FOTF can be achieved as in the standard classical integer-order case:

$$H(s)=C\left(s^{n}I-A\right)^{-1}B+D$$

3 Problem Description

A continuous-time fractional-order time-invariant system is described in (1). It is assumed the orders of differentiation are known by knowing the physics model [2] or approximating through estimating a frequency response such as in [10]. It is also considered that there is no time delay between $x(t)$ and $u(t)$. It is assumed that the differential operator is given by $p^{\alpha}=\frac{d^{\alpha}x(t)}{dt^{\alpha}}$ in (1) and their orders are commensurate. Thus (1) can be expressed as:

$$x(t)=\frac{B\left(P^{\alpha}\right)}{A\left(p^{\alpha}\right)}u(t) \tag{6}$$

where $A(p^{\alpha})=a_0p^{\alpha n}+a_1p^{\alpha n-1}+\cdots+a_n$ and $B(p^{\alpha})=b_0p^{\alpha m}+b_1p^{\alpha m-1}+\cdots+b_m$ and $x(t)$ and $u(t)$ are uniformly sampled with a sampling interval h. It yields a discrete instant $t_k=kh$ for $k=1,2,\ldots N$. The output $x(t)$ is corrupted by an additive discrete-time noise, so that the output becomes:

$$y\left(t_k\right)=x\left(t_k\right)+\varepsilon\left(t_k\right) \tag{7}$$

where $\varepsilon(t_k)$ is an autoregressive moving average (ARMA) process:

$$\varepsilon\left(t_k\right)=\frac{C\left(q^{-1}\right)}{D\left(q^{-1}\right)}e\left(t_k\right) \tag{8}$$

with $e(t)$ representing white noise.

The aim of the identification procedure is to estimate the parameters $(a_1a_2\ldots a_n)$ and $(b_0b_1\ldots b_m)$ of (6) based on the input $u(t_k)^N$ and the measured output $y(t_k)^N$ data, where the superscript N denotes N data pairs or samples.

4 Refined Instrumental Variable Continuous-Time Fractional-Order Method

The RIVC method was proposed by P.C. Young in [6] for integer-order systems. This method is extended in the paper to identify fractional-order systems. From (6) and (7), the noise process becomes:

$$\varepsilon(t_k) = y(t_k) - \frac{B(P^{\alpha})}{A(p^{\alpha})} u(t)$$

$$= A(p^{\alpha}) \frac{1}{A(p^{\alpha})} y(t_k) - B(P^{\alpha}) \frac{1}{A(p^{\alpha})} u(t)$$

Defining $y_{fA}(t_k)$ and $u_{fA}(t_k)$ as filtered forms of $y(t_k)$ and $u(t_k)$, respectively, leads to:

$$\varepsilon(t_k) = A(p^{\alpha}) y_{f_A}(t_k) - B(P^{\alpha}) u_{f_A}(t) \tag{9}$$

Consequently, $y_{fA}^{\alpha n}(t_k)$ can be expressed in regression form as:

$$y_{f_A}^{\alpha n}(t_k) = \varphi_{fA}^{T} \theta + \varepsilon(t_k) \tag{10}$$

where

$$\varphi_{f_A(t_k)} = [-y_{f_A}^{\alpha(n-1)}(t_k), -y_{f_A}^{\alpha(n-2)}(t_k), \ldots - y_{f_A}(t_k)$$
$$u_{f_A}^{\alpha(m)}(t_k), u_{f_A}^{\alpha(m-1)}(t_k), \ldots u_{f_A}(t_k)\,]^T$$

$$\theta = [\,a_1, a_2, \ldots, a_n\, b_0, b_1, \ldots, b_m\,]^T$$

$$y_{f_A}^{\alpha(n-i)}(t_k) = \frac{1}{A(p^{\alpha})} p^{\alpha(n-i)} y(t_k)$$

$$u_{f_A}^{\alpha(m-i)}(t_k) = \frac{1}{A(p^{\alpha})} p^{\alpha(m-i)} u(t_k)$$

Fig. 1 illustrates how to generate $y_{fA}^{\alpha(n-i)}(t_k)$ and $u_{fA}^{\alpha(m-i)}(t_k)$ from the sampled input and output. The analogue input and output may be obtained for example using a zero-order-hold, so that $y_{fA}^{\alpha(n-i)}(t)$ can be sampled to obtain $y_{fA}^{\alpha(n-i)}(t_k)$. The continuous-time filtered values $y_{fA}^{\alpha(n-i)}(t)$ can be expressed in a state space form directly from Fig. 1.

$$\begin{bmatrix} y_{f_A}^{\alpha(n)}(t) \\ y_{f_A}^{\alpha(n-1)}(t) \\ \vdots \\ y_{f_A}^{\alpha}(t) \end{bmatrix} = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_n \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{f_A}^{\alpha(n-1)}(t) \\ y_{f_A}^{\alpha(n-2)}(t) \\ \vdots \\ y_{f_A}(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} y(t)$$

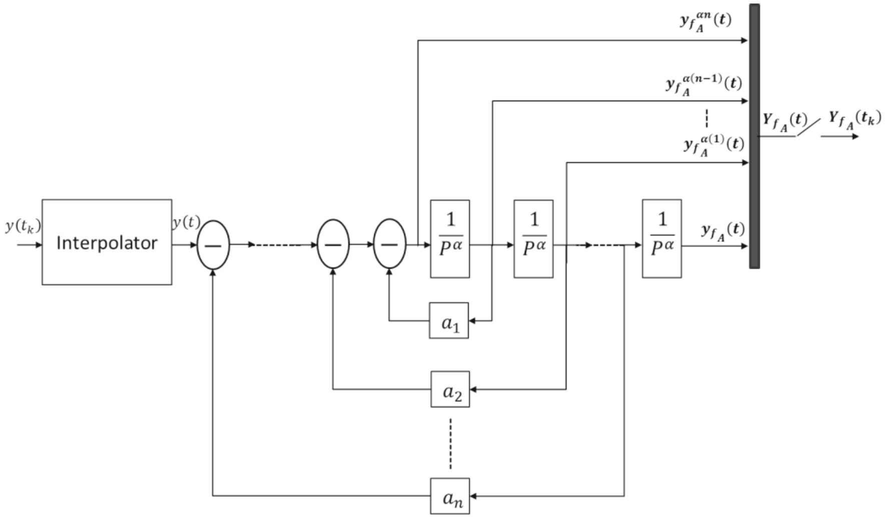


Fig. 1. Illustrates the generation of the filtered output

There are two steps in the RIVCF algorithm:

- Step 1: The initial system parameters may be obtained via various techniques such as fractional-order (FO) least squares, FO frequency analysis [10] or FO state variable filter. In this paper a use is made of the simplified least squares continuous-time fractional-order (SLSCF) method [8].
  The SLSCF method is summarised as follow:

i. Using a fractional-order commensurate stable filter whose dominator has a similar order of the $A(p^{\alpha})$ polynomial:

$$f\left(p^{\alpha}\right)=\frac{1}{\left(p^{\alpha}+\lambda\right)^{n}} \tag{11}$$

ii. In order to generate the filtered derivative of the output $y(t_k)$ and input $u(t_k)$, there is a need to pre-filter $y(t_k)$ and $u(t_k)$:

$$y_{f_f}^{\alpha(n-j)}\left(t_k\right)=\frac{1}{\left(p^{\alpha}+\lambda\right)^{n}}\,p^{\alpha(n-j)}y\left(t_k\right)$$

$$u_{f_f}^{\alpha(m-j)}\left(t_k\right)=\frac{1}{\left(p^{\alpha}+\lambda\right)^{n}}\,p^{\alpha(m-j)}u\left(t_k\right)$$

iii. The estimates can be obtained using a least squares algorithm based on (9). It is termed simplified refined least square (SRLS):

$$\hat{\theta}_1=\left(\frac{1}{N}\sum_{k=1}^{N}\varphi_{f_f}(t_k)\varphi_{f_f}^{T}(t_k)\right)^{-1}\frac{1}{N}\sum_{k=1}^{N}\varphi_{f_f}(t_k)\,y_{f_f}^{\alpha(n)}(t_k)$$

where

$$\varphi_{f_f(t_k)}=[-y_{f_f}^{\alpha(n-1)}(t_k),-y_{f_f}^{\alpha(n-2)}(t_k),\ldots-y_{f_f}(t_k)$$
$$u_{f_f}^{\alpha(m-1)}(t_k),u_{f_f}^{\alpha(m-2)}(t_k),\ldots u_{f_f}(t_k)\,]^{T}$$

iv. Making use of the estimates, repeatedly update the filter (11) by:

$$f(p)=\frac{1}{\hat{A}(p^{\alpha})}$$

(ii) to (iv) are repeated until the sum of the squares of the differences between $\hat{\theta}_i$ and $\hat{\theta}_{i-1}$ is very small.

- Step 2: The estimates are used for creating $\hat{A}(p^{\alpha})$and $\hat{B}(p^{\alpha})$. The following summarises the (RIVCF) algorithm.

  i. Based on the estimated $\hat{A}(p^{\alpha})$and $\hat{B}(p^{\alpha})$ coefficients, the noise free output can be approximated as:

$$\hat{x}(t_k)=\frac{\hat{B}(p^{\alpha})}{\hat{A}(p^{\alpha})}u(t_k) \tag{12}$$

  ii. Based on (12) the noise process can be approximated as:

$$\hat{\varepsilon}(k)=y(t_k)-\hat{x}(t_k)$$

An ARMA $\varepsilon(k)$ process in (8) can be approximated as an AR process with a much larger order of denominator [11]. Defining:

$$\frac{C(q^{-1})}{D(q^{-1})}e(k)\approx\frac{1}{\tilde{D}(q^{-1})}\hat{e}(k) \tag{13}$$

Rearranging (8) and using (13) $\hat{e}(k)$ may be approximated

$$\hat{e}(k)=A(p^{\alpha})\,y_{Df_A}(t_k)-B(P^{\alpha})u_{Df_A}(t_k) \tag{14}$$

where

$$y_{Df_A}(t_k) = \tilde{D}(q^{-1})\, y_{f_A}(t_k) \tag{15}$$

$$u_{Df_A}(t_k) = \tilde{D}(q^{-1})\, u_{f_A}(t_k)$$

The selected instrumental variable for this system is:

$$\hat{x}_{Df_A}(t_k) = \tilde{D}(q^{-1})\, \hat{x}_{f_A}(t_k)$$

iii. Equation (14) illustrates there is a need for estimating the parameters of $\tilde{D}(q^{-1})$ which can be achieved from (13). Then the filtered input $u_{f_A}(t_k)$ and output $y_{f_A}(t_k)$ can be passed through the estimated $\tilde{D}(q^{-1})$ filter in discrete-time. It is considered that $y^{\alpha n}_{Df_A}(t_k)$ is an output of the model (14). Therefore (14) can be expressed in a regression form as:

$$y^{\alpha n}_{Df_A}(t_k) = \varphi^T_{Df_A}(t_k)\,\theta + e(k) \tag{16}$$

where

$$\varphi_{Df_A(t_k)} = [-y^{\alpha(n-1)}_{Df_A}(t_k), -y^{\alpha(n-2)}_{Df_A}(t_k), \ldots - y_{Df_A}(t_k),$$
$$u^{\alpha(m-1)}_{Df_A}(t_k), u^{\alpha(m-2)}_{Df_A}(t_k), \ldots u_{Df_A}(t_k)]^T$$

iv. The estimates can be obtained using the proposal RIVCF algorithm based on (16):

$$\hat{\theta}_1 = \left(\frac{1}{N} \sum_{k=1}^{N} \hat{\varphi}_{Df_A(t_k)} \varphi^T_{Df_A}(t_k) \right)^{-1} \frac{1}{N} \sum_{k=1}^{N} \hat{\varphi}_{Df_A(t_k)} y^{\alpha n}_{Df_A}(t_k)$$

where

$$\hat{\varphi}_{Df_A(t_k)} = [-\hat{x}^{\alpha(n-1)}_{Df_A}(t_k), -\hat{x}^{\alpha(n-2)}_{Df_A}(t_k), \ldots - \hat{x}_{Df_A}(t_k),$$
$$u^{\alpha(m-1)}_{Df_A}(t_k), u^{\alpha(m-2)}_{Df_A}(t_k), \ldots u_{Df_A}(t_k)]^T$$

(ii) to (iv) are repeated until the sum of the squares of the differences between $\hat{\theta}_i$ and $\hat{\theta}_{i-1}$ is very small.

5 Numerical Example

A numerical example is presented to illustrate the performance of the RFCIV method for identification of a hybrid fractional-order Box-Jenkins model. The model is:

$$y(t_k) = \frac{2p + p^{0.5} + 1}{p^{1.5} + 2p + 3} u(t_k) + \frac{1 + 0.2q^{-1}}{1 - 0.7q^{-1}} e(k) \tag{17}$$

where $y(t_k)$ and $u(t_k)$ are the output and input, respectively. The input is selected to be a pseudo-random binary sequence with magnitude (-10, 10). The complete set of input and output contains (1000, 2000, 3000) samples with sampling interval $h = 5 * 10^{-3}$ and $e(k)$ is a white noise sequence with zero mean and 10^{-2} variance. The parameter λ in (11) is selected to be 2.

The RFCIV algorithm is applied to (17) using Monte Carlo analysis for 100 runs.

Table 1 illustrates that SRLSCF does not give accurate estimates when N=1000. However it gives better estimates as the number of samples increases. In the case of the RIVCF algorithm, Table 2 shows that much better estimates are obtained. The parameter estimation improves as the number of samples increases.

Table 1. Illustrates means and standard deviations of the five estimates of numerical example using the SRLSCF algorithm

Parameters	*N = 1000*	*N = 2000*	*N = 3000*
a_1= 2	1.1915 ± 0.0086	1.5305 ± 0.0034	1.6149 ± 0.00253
a_2= 3	1.8737 ± 0.0024	2.2835 ± 0.0094	2.4215 ± 0.0063
b_0= 2	1.9609 ± 0.0001	1.9721 ± 0.0001	1.96751 ± 0.0001
b_1= 1	-0.1298 ± 0.0173	0.3766 ± 0.0063	0.4850 ± 0.0049
b_2= 1	1.8719 ± 0.0026	0.8652 ± 0.0007	0.8861 ± 0.0005

Table 2. Illustrates means and standard deviations of the five estimates of numerical example using the RIVCF algorithm

Parameters	*N = 1000*	*N = 2000*	*N = 3000*
a_1= 2	2.0955 ± 0.1274	2.0285± 0.0105	2.0011 ± 0.0091
a_2= 3	3.1393 ± 0.2420	3.0447 ± 0.0402	2.9981 ± 0.0206
b_0= 2	2.0050 ± 0.0006	1.0375 ± 0.0001	2.0000 ± 0.0001
b_1= 1	1.1322 ± 0.2407	1.0375 ± 0.0297	1.0019 ± 0.0170
b_2= 1	1.0172 ± 0.0072	1.0086 ± 0.0020	0.9973 ± 0.0019

6 Conclusion

The paper has proposed a new algorithm specifically developed for identifying the parameters of fractional-order systems. The new algorithm is an extension of the refined instrumental variable approach developed by P.C.Young for continuous-time integer-order Box-Jenkins models.

In a similar manner the initialisation process involves a simplified version of the algorithm which does not assume measurement noise. This algorithm is used to establish the filter which is then implemented in the full hybrid algorithm.

An illustrative example has been simulated to show the performance of the proposed identification procedure.

The application of fractional-order models is wide and extends to many non-linear phenomena. Further extensions of the algorithm to handle such systems are currently under investigation.

References

1. Podlubny, I.: Fractional differential equations. Academic Press, New York (1999)
2. Das, S.: Functional Fractional Calculus for System Identification and Controls. Springer, Heidelberg (2009)
3. Oldham, K.B., Spanier, J.: The Fractional Calculus. Academic Press, San Diego (1974)
4. Young, P.C., Jakeman, A.J.: Refined instrumental variable methods of time-series analysis: Part III, extensions. International Journal of Control 31, 741–764 (1980)
5. Young, P.C.: Parameter estimation for continuous-time models-a survey. Automatica 17(1), 23–39 (1981)
6. Young, P.C., Garnier, H., Gilson, M.: An optimal instrumental variable approach for identifying hybrid continuous-time Box-Jenkins models. In: 14th IFAC Symposium on System Identification, Newcastle, Australia, pp. 225–230 (March 2006)
7. Monje, C.A., Chen, Y.Q., Vinagre, B.M., et al.: Fractional-order Systems and Controls: Fundamentals and Applications. Springer, London (2010)
8. Malti, R., Victor, S., Oustaloup, A., Garnier, H.: An optimal instrumental variable method for continuous time fractional model identification. In: Proc. of the 17th IFAC World Congress, pp. 14379–14384 (July 2008)
9. Sabatier, J., Aoun, M., Oustaloup, A., Grégoire, G., Ragot, F., Roy, P.: Fractional system identification for lead acid battery state of charge estimation. Signal Process. 86(10), 2654–2657 (2006)
10. Ghanbari, M., Haeri, M.: Order and pole locator estimation in fractional order systems using bode diagram. Signal Process. 91(2), 191–202 (2011)
11. Söderström, T., Stoica, P.: System Identification. Series in Systems and Control Engineering. Prentice Hall, Englewood Cliffs (1989)

Investigation of Model Order Reduction Techniques: A Supercapacitor Case Study

Toheed Aizad, Malgorzata Sumisławska, Othman Maganga, Oluwaleke Agbaje, Navneesh Phillip, and Keith J. Burnham

Control Theory and Applications Centre,
Coventry University, Coventry, CV1 5FB, UK

Abstract. This paper presents several different model order reduction techniques to refine an equivalent circuit high order model of a supercapacitor. The presented model order reduction techniques are: truncation based, projection based and system identification based (data based). Upon application of these techniques to the high order model, it has been found that a reduced model with sufficient accuracy can be obtained to act as a surrogate of the real system. This is evident by the ability to reduce a 60^{th} order supercapacitor model to 4^{th} order whilst preserving accuracy.

1 Introduction

Hybrid electric vehicle (HEV) powertrain architecture often requires large power transfer to and from traction motors for delivery of desirable vehicular performance. This specifically relates to regenerative braking where large amounts of current are produced during operation [1, 2]. Although, batteries currently do not possess the required power density to absorb such large current transfers, supercapacitors (SCs) are found to be suitable for this purpose. Having power densities about five to ten times greater than batteries [3] makes them ideal for intermediate power transfers. One disadvantage however is that SCs have fairly low operational voltages (1.2 - 3.5V per cell [4]) and for HEV applications they must be connected in series to provide sufficient voltage. When connected in series, the bank of SCs require cell balancing to ensure overvoltage does not occur which would lead to cell damage and performance degradation. To ensure cell balancing, models that accurately describe the dynamic behaviour of SCs are required within the control system. These models must be computationally efficient as well as being able to effectively work as surrogates for the real SC. This requirement often leads to models that are of low order but insufficiently descriptive or high in fidelity but inappropriate for online implementation. To address this need, this paper demonstrates various model order reduction techniques that successfully refine a high order model (HOM) and at the same time retain the dynamic behaviour to accurately represent a SC system for the required control frequency bandwidth.

In this paper a HOM of a SC is presented which is a candidate for model order reduction via truncation, projection and data based techniques. Each method

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_76, © Springer International Publishing Switzerland 2014

is described and compared against each other as well as against low order SC resistance-capacitance (RC) branch models for better understanding of performance and the benefits of model order reduction.

2 Supercapacitor Model

The SC is represented by an inductor L, a series resistor R_i, a complex pore impedance Z_p, and a leakage resistance R_L as shown in Fig. 1. The complex pore impedance Z_p attribute of the model relates to level of electrode porosity in the supercapacitor which directly impacts the impedance of the system. The mathematical expression for $Z_p(j\omega)$ is given by, see [5, 6]

$$Z_p(j\omega) = \frac{\tau \coth(\sqrt{j\omega\tau})}{C\sqrt{j\omega\tau}} \tag{1}$$

where τ is the time constant, C is the capacitance and *coth* is the hyperbolic cotangent.

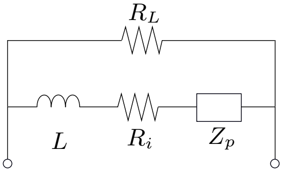


Fig. 1. Equivalent circuit of SC

It has been shown in [5] that the nonlinear complex pore impedance Z_p can be approximated by a number of RC branches in series with a capacitor, see Fig. 2, where the resistance and the capacitance of each branch are, see [5]

$$R_k = \frac{2\pi}{\pi^2 k^2 C}, \quad k = 1, 2, \cdots, n \tag{2}$$

$$C_1 = C_2 = \cdots = C_n = \frac{C}{2}, \quad n = 58 \tag{3}$$

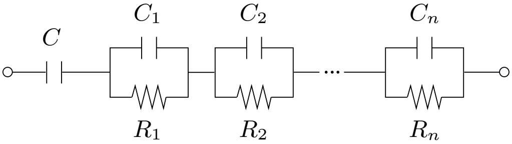


Fig. 2. Approximation of Z_p by n RC branches in series with a capacitor

Further in this paper, the SC model with Z_p given by (1) is referred to as the *analytical model*, whilst the model where the complex pore impedance is modelled by a number of RC branches is termed as the *n-branch model*.

3 Model Order Reduction Techniques

3.1 Truncation Based

Truncated Balanced Residualisation (TBR) is a popular and practical model order reduction method [6–8]. This approach was developed by Moore [9] and consists of balancing the system and then discarding the states corresponding to small Hankel singular values (HSV), see [10]. The resulting reduced order model is stable, balanced and minimal. Furthermore, the frequency error bound is easily calculable [7, 10, 11].

Consider a linear time invariant (LTI) HOM in the state-space form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times p} \\ y(t) &= Cx(t) + Du(t), \quad C \in \mathbb{R}^{p\times n}, D \in \mathbb{R}^{p\times p} \end{aligned} \tag{4}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^p$, and $y(t) \in \mathbb{R}^p$ are, respectively, the system state vector, the input, and the output. A, B, C, and D are appropriately dimensioned matrices. For computation of controllability and observability Grammians, the system must be assumed to be asymptotically stable, controllable and observable. These controllability and observability Grammians are, respectively, given by

$$\begin{aligned} W_c &= \int_0^\infty e^{At} B B^T e^{A^T t} dt \\ W_o &= \int_0^\infty e^{A^T t} C^T C e^{At} dt \end{aligned} \tag{5}$$

If the system is controllable and observable the Grammian matrices are positive definite and they satisfy the following Lyapunov equations [6, 11–13]

$$\begin{aligned} AW_c + W_c A^T + BB^T = 0 \\ A^T W_o + W_o A + C^T C = 0 \end{aligned} \tag{6}$$

The HSV of the HOM are extracted as the square roots of the products of eigenvalues of two Grammians

$$\sigma_i = \sqrt{\lambda_i(W_c W_o)}, \quad i = 1, 2, \cdots, n \tag{7}$$

such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \sigma_k \text{ and } \sigma_{k+1} = \cdots = \sigma_n \tag{8}$$

where n is the order of HOM and k is the number of non-zero HSV. The balancing transformation is the state transformation that makes the controllability and observability Grammians identical and diagonal [13], i.e.

$$W_c = W_o = \varSigma = \operatorname{diag}\left(\begin{bmatrix} \sigma_1 \; \sigma_2 \; \cdots \; \sigma_n \end{bmatrix}\right) \tag{9}$$

The two Grammians (W_c and W_o) can be factored by using Cholesky factorization to obtain balancing transformations as

$$T = L_c V \Sigma^{-\frac{1}{2}} \qquad\qquad T^{-1} = \Sigma^{-\frac{1}{2}} U^T L_c^T \tag{10}$$

where U and V are unitary matrices and L_c and L_o are lower triangular matrices. The balanced system transfer function is given by

$$G_b(s) = C_b(sI - A_b)^{-1} B_b + D \tag{11}$$

where

$$A_b = T^{-1}AT \qquad\qquad B_b = T^{-1}B \qquad\qquad C_b = CT \tag{12}$$

Let the balanced system be portioned as

$$A_b = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad B_b = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \quad C_b = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \quad D_b = D \tag{13}$$

Matrices A_{11} and A_{22} are, respectively, of dimension $r \times r$ and $(n-r) \times (n-r)$, where $r < n$ is the order of the reduced system, with the remaining matrices having dimensions consistent with the system dimension defined in (4).

The reduced order system obtained via balanced truncation is defined by

$$\begin{aligned} \dot{x}_1(t) &= A_{11}x_1(t) + B_1u(t) \\ y(t) &= C_1x_1(t) + Du(t) \end{aligned} \tag{14}$$

It is observed that the reduced order system obtained through the TBR gives a good approximation at high frequencies but displays a considerable steady state error [7, 10]. The reason for this error is the fact that the original and the reduced order system have different DC gains due to the truncation of weak modes [7, 10, 13].

Singular perturbation approximation (SPA) extends the idea of the TBR such that instead of discarding the states, their derivatives are set to zero. This preserves the DC gain of the system and retains more information about the original system than the TBR [10, 13].

Consider the balanced linear system in the form of, cf. (13)

$$\begin{aligned} \dot{x}_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + B_1u(t) \\ \dot{x}_2(t) &= A_{21}x_1(t) + A_{22}x_2(t) + B_2u(t) \\ y(t) &= C_1x_1(t) + C_2x_2(t) + Du(t) \end{aligned} \tag{15}$$

By setting the derivative of $x_2(t)$ to zero, the vector $x_2(t)$ can be expressed as

$$x_2(t) = -A_{22}^{-1}\left(A_{21}x_1(t) + B_2u(t)\right) \tag{16}$$

which leads to the reduced order system given by

$$\begin{aligned} \dot{x}_1(t) &= \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)x_1(t) + \left(B_1 - A_{12}A_{22}^{-1}B_2\right)u(t) \\ y(t) &= \left(C_1 - C_2A_{22}^{-1}A_{21}\right)x_1(t) + \left(D - C_2A_{22}^{-1}B_2\right)u(t) \end{aligned} \tag{17}$$

3.2 Projection Based

Projection based approaches for model order reduction are used to obtain reduced order models by projecting linear equations describing large scale LTI models of systems into a subspace of lower dimension [14]. The reduced order single-input single-output model being

$$\begin{aligned} \dot{x}_r(t) &= A_r x_r(t) + B_r u(t) \\ y(t) &= C_r x_r(t) + Du(t) \end{aligned} \tag{18}$$

where $x_r(t) \in \mathbb{R}^r$ is the state vector of the reduced order model, whilst matrices A_r, B_r, C_r are obtained via

$$A_r = WAV \qquad\qquad B_r = WB \qquad\qquad C_r = CV \tag{19}$$

One of such methods is the Krylov subspace based Arnoldi method which uses the Arnoldi algorithm as proposed in [15]. In order to provide projection bases $V \in \mathbb{R}^{n\times r}$ and $W \in \mathbb{R}^{n\times r}$ the Krylov subspace technique dervies the columns of V and W in such a way that a large state space $x \in R^n$ can be mapped into a smaller subspace via $x \approx Vx_r$ [16]. The projection bases used in this paper were calculated making use of the Arnoldi method which utilises a modified Gram-Schmidt orthogonalisation [11]. The Arnoldi method produces projection bases of the form of $W^T = V$, where the columns of the projection base V span the Krylov subspace K_r

$$K_r = span\{\begin{bmatrix} B & A^{-1}B & A^{-2}B & \cdots & A^{-r+1}B \end{bmatrix}\} \tag{20}$$

3.3 Data-Based

A schematic diagram of data-based model order reduction is presented in Fig. 3. Firstly, the HOM is used to generate a response to a given input signal. Subsequently, an identification technique is used to derive a low order model from known input and output data.

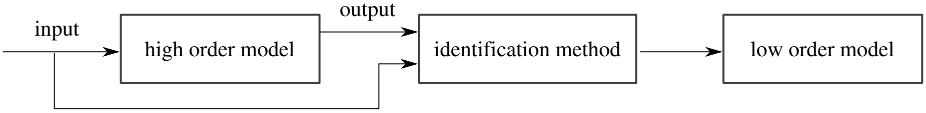


Fig. 3. Schematic diagram of data-based model order reduction

As opposed to analytical methods described in Subsections 3.1 and 3.2, the outcome of the system identification heavily depends on the quality of the data, i.e. the input and the output. The input used for the identification experiment should excite all the system modes in the considered frequency range.

Furthermore, choice of a suitable identification method is crucial. In the case of a stiff system, i.e. systems whose eigenvalues are of a different order of magnitude, continuous-time system identification is required [17]. In this paper the simplified refined instrumental variable method for continuous-time system identification (SRIVC) [18] has been applied, which produces a continuous-time model directly from the data sampled with a sampling interval of 0.5 ms.

4 Results of Model Order Reduction

4.1 Efficacy Index

In order to assess the goodness of fit of the reduced order model in the frequency domain, a quantitative efficacy index is required. In this paper the integral of absolute error (IAE) of the difference in the magnitude frequency responses between the complex and the reduced order model is used, which is calculated as

$$IAE = \int_{\bar{\omega}_1}^{\bar{\omega}_2} \left(M_{complex}(\bar{\omega}) - M_{reduced}(\bar{\omega})\right) d\bar{\omega} \tag{21}$$

where $\bar{\omega} = \log_{10}(\omega)$ and ω is the frequency expressed in radians per second. Terms $M_{complex}$ and $M_{reduced}$ refer to the magnitudes of the frequency responses of, respectively, the HOM (either the 60^{th} order n-branch or the analytical model) and the reduced order model.

4.2 Reduction of Model Order by Decreasing Number of RC Branches

As the n-branch model is an approximation of the analytical model, it is worth exploring, how well a low order n-branch model resembles the analytical model. In Fig. 4 the frequency response of the analytical model is compared with the frequency responses of n-branch models with different numbers of branches. Fig. 5 shows an increasing accuracy of the n-branch model as the model order increases.

One can note that the accuracy of the n-branch model with a low number of branches is relatively poor and it significantly increases with an increase of the number of branches. Thus, in this paper a 58-branch model (60^{th} order, see [5]) is reduced using the techniques described in Section 3 in order to obtain a relatively good approximation of the analytical model. Note the IAE between the analytical and the 58-branch model is equal to 0.14, cf. Fig. 5.

4.3 Comparison of Reduced Order Models

The state contributions of the 60^{th} order model, i.e. the Hankel singular values, are plotted in Fig. 6. One can note that the 60^{th} order model has two dominant

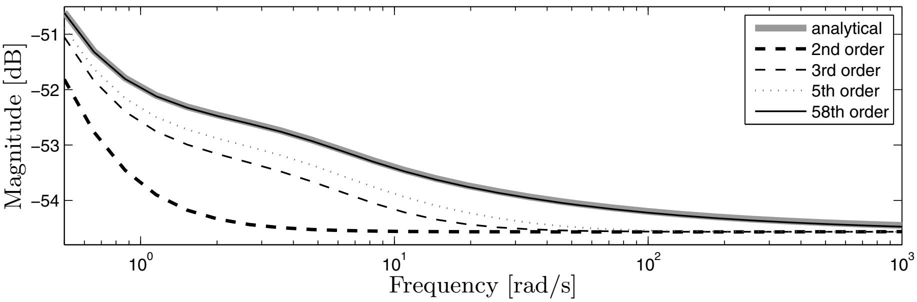


Fig. 4. Comparison of frequency responses n-branch model with analytical model of SC

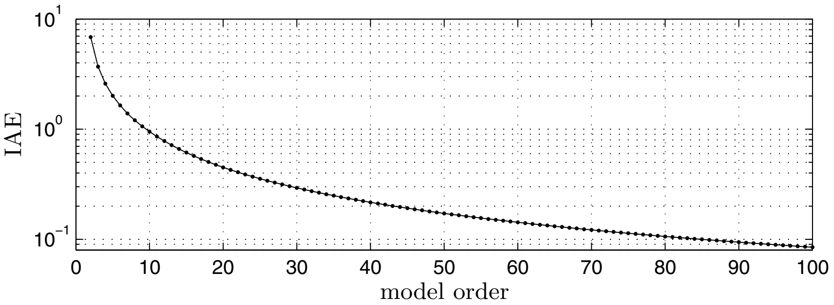


Fig. 5. Model mismatch in terms of IAE between analytical model and n-branch model of SC for different values of n (Note that $n =$ model order $- 2$.) $M_{complex}$ in (21) refers to frequency response of analytical model, whilst M_{model} refers to n-branch model

modes with the highest contribution; thus, the second order model has been selected as a starting point for the model order reduction process.

Fig. 7 compares frequency responses of second order reduced models of the SC obtained using different techniques. Both TBR and SRIVC provide relatively good match compared to the SPA and the Arnoldi method. An improvement of model accuracy has been achieved by increasing the order of the reduced model, see Fig. 8 and Fig. 9.

In Table 1 values of the IAE for reduced order models obtained using different techniques are given. It is noted that the low (3^{rd}-5^{th}) order model obtained by reduction of the 60^{th} order model has the accuracy comparable to the 58-branch model in terms of the IAE.

Accuracy of both truncation-based and projection-based techniques increases with an increase of the order of the reduced model. However, this is not always the case when the data-based SRIVC is used. An increase of the order of the model identified from the data may lead to an overparameterisation of the model which makes the identification algorithm more susceptible to errors/uncertainties.

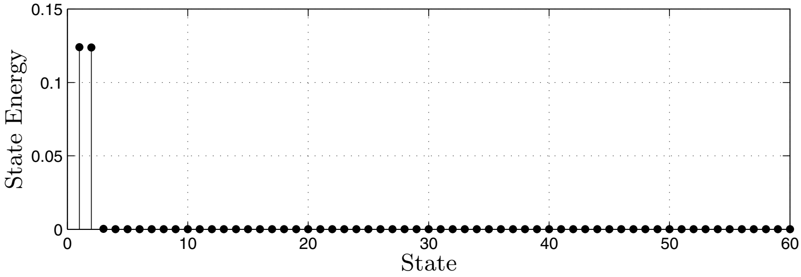


Fig. 6. Hankel singular values (state contribitions) of 58-branch model

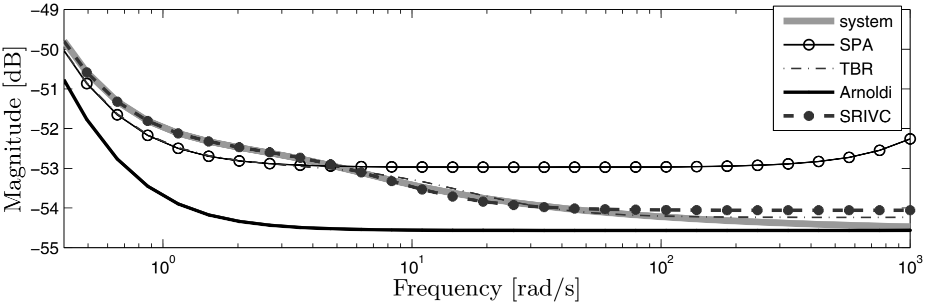


Fig. 7. Frequency responses of second order models obtained using various techniques

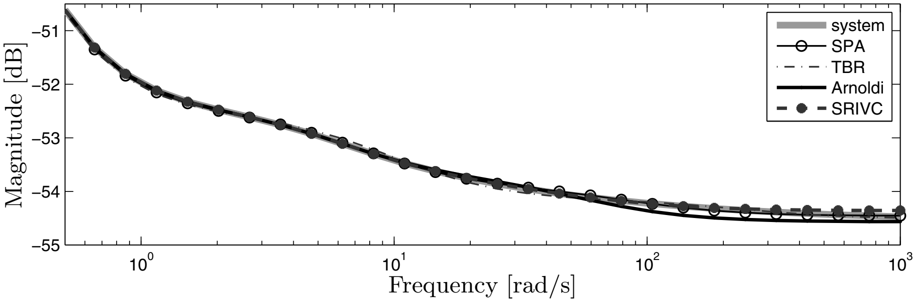


Fig. 8. Frequency responses of reduced, 4^{th} order models

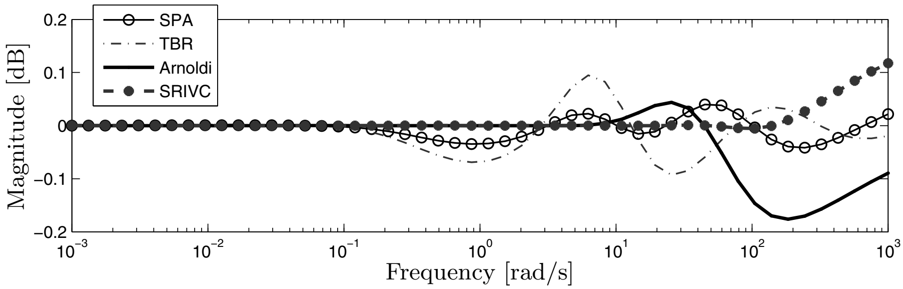


Fig. 9. Differences in magnitude frequency responses between the 60^{th} order model and reduced 4^{th} order models

Table 1. The IAE of reduced order models. Numbers outside brackets in columns 2-5 present the IAE between the 58-branch model and reduced order models; numbers in brackets refer to the IAE between the analytical model and the reduced order models. The last column presents the IAE between the analytical model and the appropriate low order n-branch model (with $n = 0, 1, 2, 3$). The IAE has been calculated for $\omega_1 = 0.001$ rad/s, $\omega_2 = 1000$ rad/s

r	TBR	SPA	Arnoldi	SRIVC	n-branch model
2	1.19 (1.19)	5.90 (5.84)	6.72 (6.86)	1.06 (1.04)	6.86
3	1.19 (1.26)	0.60 (0.60)	1.19 (1.27)	0.17 (0.23)	3.70
4	0.33 (0.39)	0.16 (0.23)	0.38 (0.47)	0.10 (0.17)	2.59
5	0.13 (0.20)	0.04 (0.14)	0.13 (0.23)	0.34 (0.41)	2.01

5 Conclusions

Model order reduction of a high order SC model is considered. It has been shown that truncation based, projection based, and data based techniques are suitable to produce simplified model formulations while preserving acceptable accuracy. It has been observed that by approximating a high fidelity nonlinear analytical model of SC by a high order equivalent circuit linear model and then reducing the order of the linear model one can obtain a 4^{th} order model, which is suitable for control (SC cell balancing) yet its accuracy is comparable to the accuracy of the high order linear model.

Acknowledgement. This work is part of the EPSRC funded FUTURE Vehicles project (EP/I038586/1). The authors also acknowledge the support from project partners.

References

1. Ashtiani, C., Wright, R., Hunt, G.: Ultracapacitors for automotive applications. Journal of Power Sources 154, 561–566 (2006)
2. Dixon, J.W., Ortuzar, M.E.: Ultracapacitors + DC-DC converters in regenerative braking system. IEEE Aerospace and Electronic Systems Magazine 17, 16–21 (2002)
3. Miller, J.R., Burke, A.F.: Electrochemical capacitors: Challenges and opportunities for real-world applications. Electrochemical Society Interface, 53–57 (2008)
4. Zhang, J., Zhang, L., Liu, H., Sun, A., Liu, R.: Electrochemical Technologies for Energy Storage and Conversion. John Wiley & Sons (2011)
5. Buller, S., Karden, E., Kok, D., de Doncker, R.: Modeling the dynamic behaviour of supercapcitors using impedance spectroscopy. IEEE Transactions on Industry Applications 38(6) (2002)
6. Cingoz, F., Bidram, A., Davoudi, A.: Reduced order, high-fidelity modeling of energy storage units in vehicular power systems. In: Proceedings of the IEEE Vehicle Power and Propulsion Conference (2011)
7. Liu, Y., Anderson, B.: Singular perturbation approximation of balanced systems. International Journal of Control 50(4), 1379–1405 (1989)
8. Wang, S., Wang, B.: An example of balanced truncation method and its surprising time domain performance. In: IEEE Conference on Cybernetics and Intelligent Systems (2008)
9. Moore, B.: Principal component analysis in linear system: controllability observability and model reduction. IEEE Trans. Automat. Contr. AC 26, 17–32 (1981)
10. Samar, R., Postlethwaite, I., Gu, D.W.: Applications of the singular perturbation approximation of balanced systems. In: Proceedings of the Third IEEE Conference on Control Applications (1994)
11. Tan, D., He, L.: Advanced model order reduction techniques in VLSI design. Cambridge University Press (2007)
12. Kumar, D., Tiwari, J., Nagar, S.: Reduction of large scale systems by extended balanced truncation approach. International Journal of Engineering Science 3(4), 2746–2752 (2011)
13. Gajic, Z., Lelic, M.: Singular perturbation analysis of system order reduction via system balancing. In: Proceedings of the America Control Conference (2000)
14. Phillips, J.R.: Projection frameworks for model reduction of weakly nonlinear systems. In: Proceedings of the 37th Annual Design Automation Conference (2000)
15. Arnoldi, W.E.: The principle of minimized iterations in the solution of the matrix eigenvalue problem. Quart. Appl. Math. 9(1), 17–29 (1951)
16. Dong, N., Roychowdhury, J.: Piecewise polynomial nonlinear model reduction. In: Proceedings of Design Automation Conference (2003)
17. Garnier, H., Wang, L. (eds.): Identification of Continuous-time Models from Sampled Data. Springer (2008)
18. Young, P.C.: Recursive Estimation and Time Series Analysis: An Introduction for the Student and Practitioner. Springer (2011)

Erratum: Matlab Simulation of Photon Propagation in Three-Layer Tissue

Julia Kurnatova[1], Dominika Jurovata[1], Pavel Vazan[1], and Peter Husar[2]

[1] Slovak University of Technology in Bratislava, Faculty of Materials Science
and Technology in Trnava, Hajdóczyho 1, Trnava 917 24, Slovak Republic
[2] Ilmenau University of Technology, Institute of Biomedical Engineering
and Informatics, 98693 Ilmenau, Germany

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_40, © Springer International Publishing Switzerland 2014

DOI 10.1007/978-3-319-01857-7_77

The paper "Matlab Simulation of Photon Propagation in Three-Layer Tissue" by Julia Kurnatova, Dominika Jurovata, Pavel Vazan, Peter Husar, DOI 10.1007/978-3-319-01857-7_40, appearing on pages 415-423 of this volume has been retracted due to a serious case of plagiarism. It is a plagiarized version of the paper "Simulation of Photon Propagation in Tissue Using Matlab" by Dominika Jurovata, Julia Kurnatova, Sebastian Ley, Daniel Laqua, Pavel Vazan, Peter Husar.

The original online version for this chapter can be found at
http://dx.doi.org/10.1007/978-3-319-01857-7_40

Author Index